Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

10-21-2020 11:00 AM

Cancer Detection in Radical Prostatectomy Histology using Convolutional Neural Networks

Laurie Huang, *The University of Western Ontario*

Supervisor: Ward, Aaron D., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Medical Biophysics
© Laurie Huang 2020

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Translational Medical Research Commons

Recommended Citation
Huang, Laurie, "Cancer Detection in Radical Prostatectomy Histology using Convolutional Neural Networks" (2020). *Electronic Thesis and Dissertation Repository*. 7388.
https://ir.lib.uwo.ca/etd/7388

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Radical prostatectomy (RP) is a common treatment for prostate cancer. We used RP whole-mount tissue samples from 68 patients stained with haematoxylin and eosin to create cancer maps using four pretrained networks: AlexNet, NASNet, VGG16 and Xception, to classify regions of interest (ROIs) as cancer or non-cancer. Models were trained on either raw images or as tissue component maps (TCMs) containing nuclei, lumina and stroma/other components generated from a trained U-Net.

All models performed similarly; however, VGG16 trained on raw images performed with the highest area under the receiver operating characteristic curve (AUC) of 0.994 (95% confidence interval 0.992-0.996). Ensemble models using models trained on raw images performed with the lowest false positive rates. All models had high false negative rate for Gleason 5 cancer and those trained on raw images performed with higher AUC than models trained on TCMs.

Keywords: Prostate cancer, radical prostatectomy, histopathology, machine learning, adenocarcinoma, cancer detection

Summary for Lay Audience

Prostate cancer is the second most frequently diagnosed cancer in men worldwide. Although this is a very treatable disease, survival rates of those with aggressive cancer is much lower than those with a less aggressive form. One treatment option available for those with gland localized cancer is radical prostatectomy, where the entire gland is surgically removed.

Expert contouring and grading of the aggressiveness of cancer on tissue slides could provide valuable information on patient prognosis, as well as guide treatment plans after the surgery. Unfortunately, detailed contours are incredibly time consuming and impractical to do in a clinical setting, so there is an unmet need for an automated cancer contouring algorithm. As a part of tissue processing, dyes are applied so the tissue is visible under a microscope. Application of the dye and the tissue properties itself can affect how intensely the dye appears, making it difficult for algorithms to be robust to the colour intensities.

In this thesis we use deep learning for cancer detection. Deep learning is a technique that teaches an algorithm to find patterns in a set of data and is a relatively young field with promising results in computer vision. Deep learning has been used to detect cancer in radical prostatectomy histology, but the papers we surveyed had a limited number of patients, or only tested one deep learning model. That is why we are comparing four deep learning models in cancer detection.

We used these four deep learning models to generate cancer maps from tissue slides and found that these models perform comparatively to each other. We produced accurate cancer maps but further research is needed to analyze potential impact on the clinical and research workflow.

Co-Authorship

This thesis is written in an integrated article format.

Chapter 2: L. Huang, W. Han, J.A. Goméz, M. Moussa, S.E. Pautler, J.L. Chin, G.S. Bauman and A.D. Ward are authors for the article ”Cancer detection on digitized prostatectomy slides using convolutional neural networks” which is currently in preparation for submission. My work on the chapter includes: defining the research question, formatting the experimental design, writing the program code, the statistical analysis and drafting the manuscript. A.D. Ward who was the supervisor of this project, contributed to defining the research question and formatting the experimental design. G.S. Bauman was the principal investigator of the study where the specimens were obtained. J.L. Chin and S.E. Pautler recruited patients and performed prostatectomies from which the histology slides were collected. M. Gaed contoured and graded the histology slides. M. Moussa and J.A. Goméz verified the contours and annotations. W. Han contributed to defining the research question and to digitally preprocessing histology slides.

Acknowledgements

This year has certainly been a challenge given everything that has happened and the massive uncertainty everyone in the world has faced. I feel that it is important, more than ever, that I show my gratitude and appreciation for those who have helped me in my time at Western University.

Firstly, I would like to thank my supervisor Aaron Ward. Without his guidance and knowledge, this project would not be where it is. I am grateful for his patience and I truly could not have asked for a better supervisor.

I would like to thank my advisory committee Ali Khan, and Jose Goméz, as well as Charlie McKenzie for their help and taking time out of their day to provide feedback. Their insight has helped me think critically about the project and push it to new directions.

I would like to express my gratitude to my colleagues Salma Dammak, Ryan Alfano, Chris Smith, Carol Johnson, David DeVries, Andrew Warner, David Palma, as well as everyone else in the Baines Imaging Laboratory and Victoria Hospital. They made working at the lab a lot more fun. I would like to thank Wenchao Han for his mentorship and advice, without it the project would not be where it is. I also extend my thanks to my classmates and everyone in the Medical Biophysics Department. They have made my time at Western a lot more enjoyable.

Finally, I would also like to acknowledge the support my friends and family have given me over the years. It has been a long journey to get to where I am today and I could not have done it without them.

Contents

Certificate of Examination **ii**

Abstract **ii**

Summary for Lay Audience **iii**

Co-Authorship **iv**

Acknowledgements **v**

List of Figures **ix**

List of Tables **xii**

List of Appendices **xiii**

1 General Introduction **1**

- 1.1 Introduction 1
- 1.2 Background 2
  - 1.2.1 Radical Prostatectomy 2
  - 1.2.2 Adjuvant and Salvage Radiotherapy 2
  - 1.2.3 Histology 3
  - 1.2.4 Staining Variability 3
  - 1.2.5 Tissue Component Maps 4
  - 1.2.6 Gleason Grading 5

1.2.7 Observer Variability 6
1.2.8 Deep Learning 8
1.2.9 U-Net 9
1.3 Previous work 10
1.3.1 Automated Cancer Detection Algorithms 10
1.4 Thesis Outline 12

2 Cancer Detection on Digitized Prostatectomy Slides using Convolutional Neural Networks 22

2.1 Introduction 23
2.2 Methods 25
2.2.1 Materials 25
2.2.2 Network Architecture 27
2.2.3 Training 28
2.3 Results 29
2.3.1 Tissue Component Maps 29
2.3.2 Cancer Detection 32
2.4 Discussion 34
2.4.1 U-Net Segmentations 34
2.4.2 Cancer Detection 40
2.4.3 Limitations 46
2.5 Conclusions 46

3 General Discussion and Conclusions 52

3.1 Contributions of the Thesis 52
3.2 Limitations 53
3.3 Applications and Future Directions 54

4 Table of Abbreviations and Symbols 59

A Permissions to Reproduce Previously Published Material **61**

Curriculum Vitae **63**

List of Figures

1.1 H&E stained histology from whole-mount RP slides, showing the variation in staining intensity. Slides were taken from different patients and at the same centre. 4

1.2 Raw histology and corresponding tissue component map where red is nuclei, blue is lumen and green is all other tissue. 5

1.3 Gleason patterns from the 2005 ISUP update [23]. Reproduced with permission. 7

1.4 U-Net architecture used in the thesis. The model has been modified from the original to accept an input image size of 240 × 240 pixel images and output a segmentation of the same size. Modified from Ronneberger et al. [35]. 10

2.1 Full resolution and down sampled images using nearest neighbour approximation where the resolution of (a) is 0.5 μm/px and (b) is 2 μm/px. Both images correspond to 480 μm by 480 μm area of tissue. 26

2.2 TCM generated using U-Nets (e), compared against TCMs generated using adaptive thresholding (b). In (b) and (e) red indicates nuclei, blue lumen and green other. In (c) and (d) black indicates background and white indicates the corresponding tissue. 30

2.3 Error metrics for nuclei and lumen segmentation using 2-fold U-Net on either the TCIA or IGPC data in a boxplot where error bars represent 95% CI. “*” indicates a p-value less than 0.05 using a 2-tailed Mann-Whitney-Wilcoxon test. 32

2.4 Error metrics of cancer classification from a 2-fold cross validation scheme."*" indicates a p-value less than 0.05 using a 2-tailed Mann-Whitney-Wilcoxon test. Error metrics were calculated on a per-patient basis. Error bars indicate 95% CI. 33

2.5 Error maps of nuclei and lumen generated from U-Net and adaptive thresholding where white is true positive, red is false positive, black is true negative and blue is false negative. The segmentations were compared against manual thresholding. 35

2.6 Error maps of nuclei and lumen generated from U-Net and adaptive thresholding where white is true positive, red is false positive, black is true negative and blue is false negative, for a lightly stained example. The segmentations were compared against manual thresholding. 36

2.7 TCM generated using U-Net, compared against TCMs generated using manual thresholding for TCIA data. In (b) and (c) red indicates nuclei, blue lumen and green other in the TCMs. In (d) and (e) white is true positive, red is false positive, black is true negative and blue is false negative. The resolution of these images was not reported in the TCIA database. 39

2.8 ROC plots with AUC for patient with G5 cancer in Figure.2.9 42

2.9 Whole mount histology slide cancer maps. The cancer maps were produced from AlexNet, NASNet, VGG16, or Xception trained on either raw images or TCM. (a) is the WSI raw slide, (b) is the grade legend for the WSI and (c)-(m) are cancer maps where red is false positive, cyan is false negative, black is true positive and grey is true negative. 43

2.10 Whole mount histology slide cancer maps. The cancer maps were produced from AlexNet, NASNet, VGG16, or Xception trained on either raw images or TCM. (a) is the WSI raw slide, (b) is the grade legend for the WSI and (c)-(m) are cancer maps where red is false positive, cyan is false negative, black is true positive and grey is true negative. 44
2.11 FNR of different cancer grades. Error bars indicate 95% CI. No error bars were given for G5, G5+4 and G4+5 because labels were taken from only one patient. 45

List of Tables

1.1 Cancer classification studies. Acc. indicates accuracy and LOPO indicates leave one patient out. Tissue processed area is estimated assuming 1,200 mm^2 per RP slide, 10 mm^2 per biopsy and 0.55 mm^2 per TMA sample. 13

2.1 Model hyperparameters. Lr indicates learning rate and BC binary crossentropy. Adam refers to the adaptive moment optimizer [22]. 29

2.2 Segmentation mean and 95% CI of TCMs generated with a U-Net, validated against validation ROIs segmentations generated by manual thresholding. 136 ROIs from TCIA and 220 from TCIA were used for validation. 31

2.3 Cancer detection error metrics mean and 95% CI which were calculated on a per patient basis. Bold-face indicates the measurement with the lowest mean of the columns FNR, FPR and Error Rate and the highest for AUC. 34

List of Appendices

Appendix A: Permissions 61

Chapter 1

General Introduction

1.1 Introduction

Prostate cancer (PCa) is one of the most common cancers in Canada, with 1 in 9 Canadian men expected to be diagnosed in their lifetime [1]. It is a variable disease, where 5-year net survival is 98% when accounting for all stages, but only 31% when counting metastatic cancer [2]. Although that is the case, PCa is often considered over-treated [3], so it is important to identify patients who would benefit the most from a given treatment.

Radical prostatectomy (RP) is a common procedure, where the prostate gland is surgically removed. Pathology reports on RP are mostly qualitative and are important in predicting post-surgery outcome through denoting cancer grading, positive margins and diagnosis of extraprostatic extension (EPE). Currently these reports do not include annotated histology slides. Detailed manual tumour annotations of whole slide images (WSIs) are currently too labour intensive to be implemented clinically, but could provide important pathological information for post-operative targeted radiotherapy [4]. An automated or assistive contouring program may be useful.

Deep neural networks (DNNs) are a form of artificial intelligence that have been used for cancer detection in RP WSIs [5, 6, 7], but these studies only test one DNN. In this thesis we use four DNNs to produce cancer maps from RP WSIs and compare their error metrics.

1.2 Background

1.2.1 Radical Prostatectomy

RP is generally done for those with suspected gland localized cancer and a life expectancy ≥ 10 years [8]. Biochemical recurrence-free survival rates are approximately 70% within 10 years after surgery [9, 10]. Complications include wound infection, erectile dysfunction and urinary incontinence [11].

1.2.2 Adjuvant and Salvage Radiotherapy

After surgery, patients may undergo adjuvant radiotherapy where they receive radiotherapy immediately to reduce the risk of recurrence. Alternatively an active surveillance approach, where the patient is monitored for biochemical recurrence and treatment is given when appropriate, may be suggested. If radiotherapy is only done after biochemical recurrence, it is called salvage radiotherapy.

Complications related to radiotherapy include gastrointestinal and genitourinary toxicity [12]. Rates of urinary incontinence are generally similar for those who have and have not received adjuvant radiotherapy, and the potential development of secondary malignancies post-RP are inconclusive [12].

Numerous studies compare patient outcomes between adjuvant and salvage radiotherapy. A meta-analysis done by Morgan et al. of patients post RP with either cancer extending beyond the prostatic capsule or in the resection margins, found adjuvant radiotherapy had not shown improved overall survival when comparing to active surveillance, but shown to increase biochemical progression-free survival [13]. In contrast, Wiegel et al. found that progression free survival at 10 years was 56% for patients with adjuvant therapy and 35% for patients with active surveillance, where progression is defined as biochemical recurrence, clinical recurrence or death [14]. Guidelines from the American Urological Association (AUA) and the American Society for Radiation Oncology (ASTRO) in 2013, revised in 2019, concluded that given a literature search of 48 adjuvant and 137 salvage radiotherapy studies, it was not possible to determine which treatment is superior due to different radiotherapy protocols, patient stratification, definitions of failure and other factors across trials [12, 10].

1.2.3 Histology

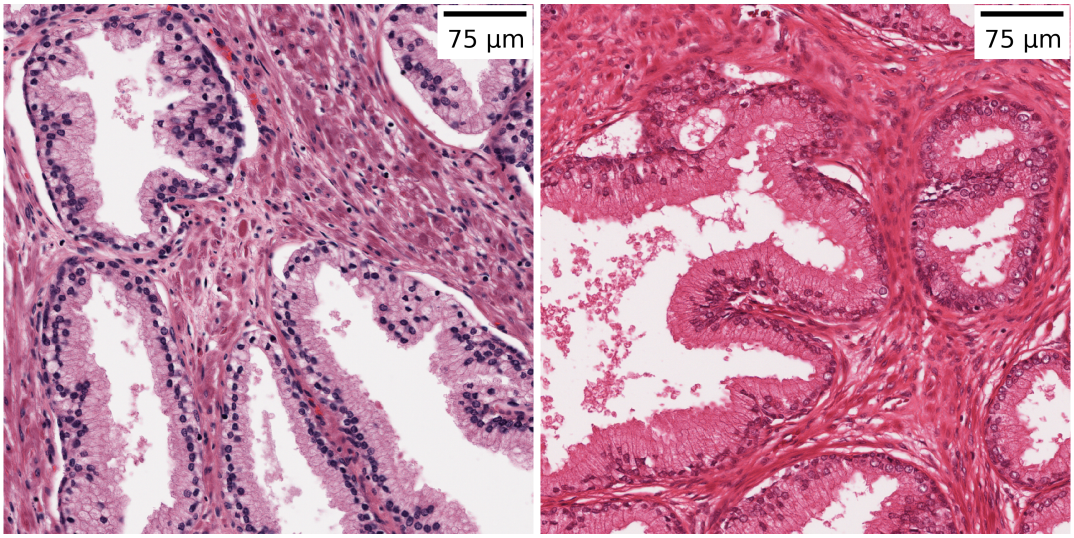
After RP, the gland is processed into tissue slides. The process involves formalin fixation and removal of the apex and base. The pieces are then embedded in paraffin and are processed either as whole-mount or standard sections where the tissue is cut into quarters.

The tissue is then stained with haematoxylin and eosin (H&E), where haematoxylin stains cell nuclei blue or purple and eosin stains the cytoplasm and extracellular matrix as pink. The stained tissues are then mounted on slides and digitally scanned with a resolution dependent on the digital scanner.

1.2.4 Staining Variability

Staining variation comes from a multitude of factors like specimen preparation, temperature and age of solutions, slice thickness, scanner variation and inter-patient variation [15]. These factors lead to variability in the colour intensity and saturation which make comparison between slides difficult—especially across centres.

Figure 1.1 shows staining variability between manually stained, whole-mount, whole slide images (WSIs) collected from two same centre RP patients. Nuclei in Figure 1.1a is more blue/purple than Figure 1.1b, which is more pink/red and is closer in colour to the cytoplasm. Although nuclei are still identifiable with the human eye in Figure 1.1b, if a computer algorithm was calibrated to segment nuclei on tissue with the staining intensity found in Figure 1.1a, the



(a) Patient A (b) Patient B

Figure 1.1: H&E stained histology from whole-mount RP slides, showing the variation in staining intensity. Slides were taken from different patients and at the same centre.

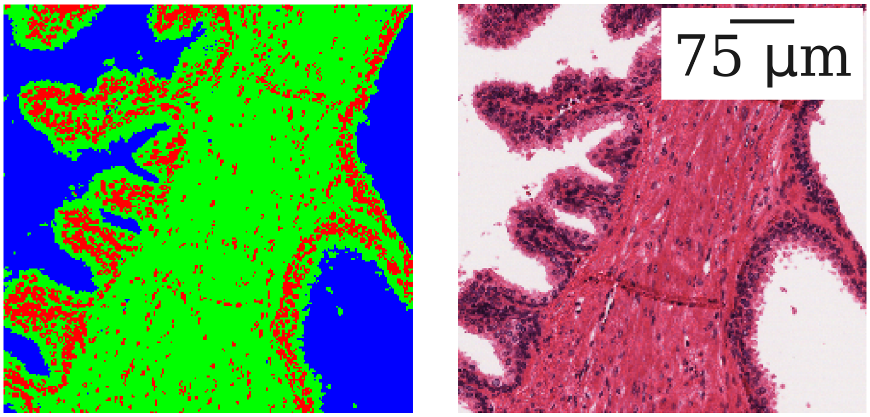
algorithm may not be able to segment differently stained tissue. Since staining intensity is sensitive to small variations in the pathology process, there have been multiple studies done to account for this variation.

Some of the studies use a colour correcting algorithm to standardize the staining intensity. Many of these algorithms rely on stain deconvolution which requires prior knowledge of stain vectors [16]. This includes a research paper published from our laboratory, which uses a stain deconvolution algorithm and adaptive thresholding [17]. Manual approaches estimate the stain vectors through a selection of representative sample pixels [16]. Unfortunately, when scaled to large studies involving hundreds of slides, manual estimation is impractical. Some studies account for this by using automatic stain vector generators [18, 19, 20].

1.2.5 Tissue Component Maps

Tissue component maps (TCMs) are histology slides broken down into its tissue composition. For example, a TCM containing nuclei, lumen and all other tissue will be a three colour map where each colour represents either nuclei, lumen or other. An example can be found in Figure

1.2. If we can create a TCM segmentation algorithm that is invariant to staining intensity, algorithms that use the TCMs do not have to adjust for staining variability.



(a) Tissue component map (b) Raw Histology

Figure 1.2: Raw histology and corresponding tissue component map where red is nuclei, blue is lumen and green is all other tissue.

TCMs are a simplification of the original histology. When it comes to cancer detection, simplification of raw histology into tissue components has been shown to produce accurate results. Kwak et al. found that best performing models for cancer detection mostly used nuclei seed maps instead of RGB raw images [21]. Han et al. also found that deep networks trained on TCMs perform with a higher area under the receiver operator curve (AUC) than raw images [7]. By removing confounding information found in the tissue, cancer classification using TCMs may perform with higher accuracy than those using raw images.

1.2.6 Gleason Grading

Tumour grading is done with the Gleason grading system. This method originated in the 1960s before it was updated by the International Society of Urological Pathology (ISUP) in 2005 and again in 2014 [22, 23, 24].

The Gleason grading system labels tissues into one of five grades from 1-5 based on tissue pattern. Gleason 1 is defined by well differentiated and uniform glands while Gleason pattern 5 is defined by poorly-differentiated glands [23]. The system was updated in 2005 to better reflect treatment changes and a growing understanding of the disease in the years since the original grading scheme was created [23]. Examples of the growth patterns of each grade from the 2005 update is shown in Figure 1.3.

Along with the Gleason grades was the Gleason Score (GS). Tissues were given two grades, indicating the primary and secondary grade. These two values were added to create the GS. For example, if a tissue sample was predominantly G3 and the secondary pattern is G4 then the Gleason score would be G3+G4 = GS7. If only one pattern was found, then the primary and secondary grades would be considered identical. For example, a tissue with only Gleason pattern 3 would have a Gleason score G3+G3 = GS6.

In the ISUP 2014 update, the grade groups were redivided into five. Grade group 1 is for Gleason scores less than or equal to 6, grade group 2 is for G3+G4=GS7, grade group 3 is for G4+G3 = GS7, grade group 4 is G4+G4=G8 and grade group 5 is GS9-10 [24]. One of the most important updates in ISUP 2014 redefined grade pattern 3 to not include cribriform pattern. These updates were done to more accurately stratify tumours, decrease the number of grading categories and reduce over treatment which had become a concern [24].

Digitally grading and contouring the slide is a very time intensive process. In our lab, it took an average of 70 hours for a physician to digitally contour and grade mid-gland whole mount slides for a patient. The long time frame makes detailed contours impractical for clinical use, but localizing where the cancer was may help for decisions on where to target radiotherapy in the prostate bed.

1.2.7 Observer Variability

Pathologist reports, like anything that is reliant on human interpretation, are vulnerable to inter and intraobserver variability. This can be compounded by staining variability. The 2005 and 2014 ISUP updates were partially done to reduce variability in Gleason grading.

Before the 2005 ISUP update, it was found that general pathologists had more disagreements in Gleason scores than urological pathologists and a tendency to overgrade [25, 26, 27].

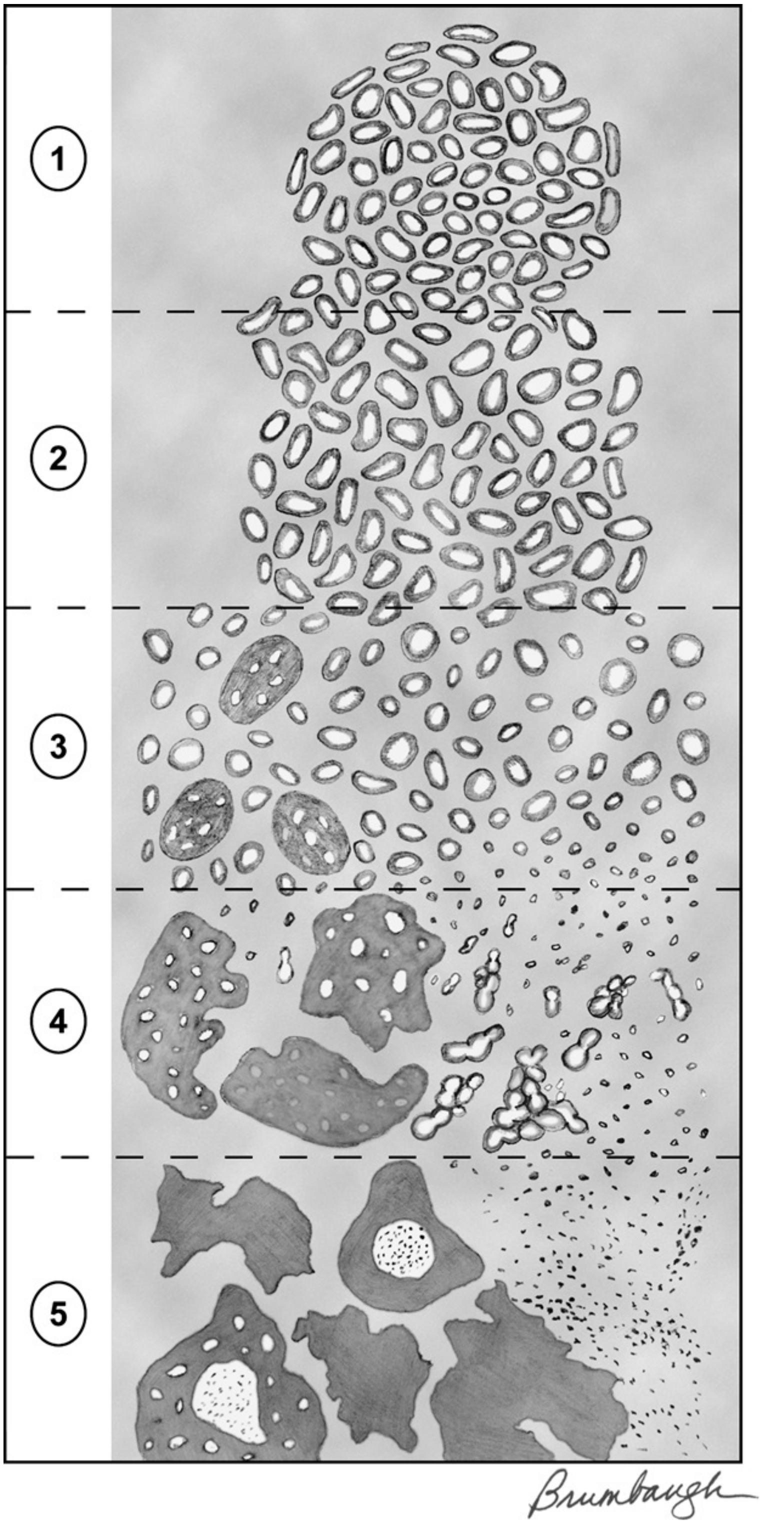


Figure 1.3: Gleason patterns from the 2005 ISUP update [23]. Reproduced with permission.

One study found intraobserver reproducibility to range from 65% to 100% [28]. After the 2014 ISUP update, interobserver agreement of Gleason grade groups using the 2014 ISUP guidelines was 51.7% between two pathologists from biopsy cores [29].

Aside from Gleason scores, several studies have found notable interobserver variability in EPE diagnosis between experienced and non-expert reviewers [30, 26]. Evans et al. also studied interobserver variability in EPE in RP and concluded that interobserver variability was related to the lack of clearly definable prostatic capsule [31].

An automated cancer detection algorithm may help the consistency of pathology grading, for either inter and intraobserver variability. A recent paper has also shown that artificial intelligence (AI) assistive program increased agreement between a panel of pathologists for Gleason grading of prostate biopsies[32].

1.2.8 Deep Learning

Deep learning is a subset of the machine learning family. Origins of deep learning are often attributed to Ivakhnenko and Lapa in their work featuring multilayer perceptrons in 1967 [33]. It was not until the development of graphics processing units (GPUs) in the early 2000s, did advancement in DNN research accelerate. This meant that more complex deep learning models can be trained in a feasible time span.

DNNs essentially create predictions based on a set of data called the “training” data, so the quality of the model and data are closely linked. In classification problems, data are labelled and the goal of a DNN is to predict the label correctly. When the model outputs an incorrect classification, the model adjusts its parameters accordingly. If the model predicts correctly, then it does not change. This way the model “penalizes” incorrect predictions and “rewards” correct ones. A “validation” dataset, a dataset which has not been used for training, may be used to assess how the model is preforming while it is training. After multiple iterations the model eventually converges to a solution that may or may not perform at the desired accuracy.

The quality of the model can be assessed after training through the performance on a set of unseen data known as the “testing” data. Testing data must be kept separate from the training and validation data to accurately measure the model performance and prevent bias. If the training data was a poor representation of the population or had a lot of noise, then the model may not be able to classify new data as it did not converge to a solution effective on both datasets. It is for that reason DNN performances are highly dependent on its training data.

The large dataset generally required for deep learning makes applications in the medical field difficult due to privacy concerns and expertise required to label data. Some innovations have been made to decrease the required data size and training time. One of these techniques is “transfer learning” where a model is trained on one data set for one purpose, then those parameters are used as the initial parameters before training for another problem. Transfer learning is based on the theory that there are universal patterns used to identify images, regardless of the problem. For example, detecting edges in an image may be important for both facial and nuclei detection. By using a pretrained model, the model might be closer to converging to a solution than if the model was initialized randomly. A very common database used for pretraining is the open access ImageNet database that includes 14 million images of common objects [34].

1.2.9 U-Net

A very popular segmentation DNN model in medical imaging is the U-Net [35]. The U-Net uses a series of deep learning layers that down-sample the input image, before it is up-sampled to an output segmentation. The design resembles a “U” hence the name “U-Net”. It won the International Symposium on Biomedical Imaging (ISBI) cell tracking competition in 2015 for accurate and fast segmentations. In this thesis, we used the U-Net as a semantic segmentation algorithm on histology to produce either lumen or nuclei maps. Figure 1.4 is a diagram of the modified U-Net used in the thesis.

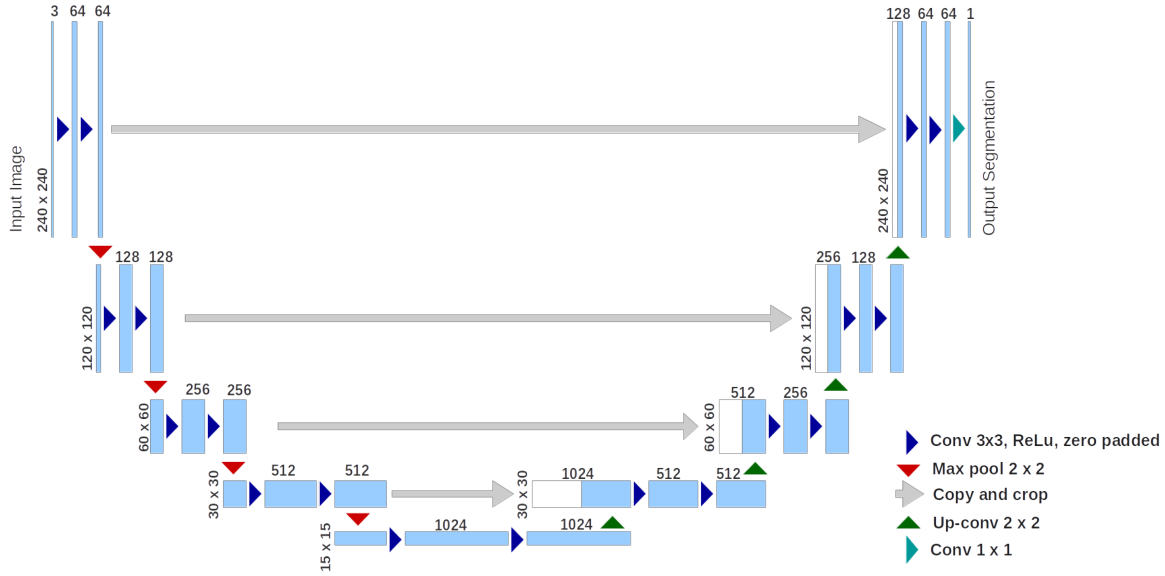


Figure 1.4: U-Net architecture used in the thesis. The model has been modified from the original to accept an input image size of 240 × 240 pixel images and output a segmentation of the same size. Modified from Ronneberger et al. [35].

1.3 Previous work

1.3.1 Automated Cancer Detection Algorithms

There have been many studies that classify prostate histology as cancer or non-cancer. These studies were reviewed and summarized in Table 1.1. Many of these studies detect cancer in biopsy slides [36, 37, 21, 38, 39].

Due to the size of a biopsy core from a 18 gauge needle (1.27 mm outer diameter), many of these studies process a smaller tissue area in comparison to those involving RPs. Biopsies also have a positive sampling bias and serve a different purpose than RP. RP histology samples the entire cross section of the prostate gland, while biopsies are diagnostic and sample a small subsection of tissue.

Bulton et al. and Ström et al. train models for cancer diagnosis using several thousand biopsy samples, resulting in a total processed tissue area comparable or greater than studies involving RP [38, 39]. These two studies detect and grade cancer on a per biopsy basis rather than per region. In machine learning these problems are called multiple-instance-learning (MIL) where the model must label a set of data, rather than each individual data point in a set of data.

Kwak et al. classified tissue micro arrays (TMAs) from biopsies as cancer or non-cancer using nuclear seed maps [21]. Computation time was 10 minutes per sample, where majority of the time is allocated to the production of nuclear seed maps. Given that one sample has an area of approximately 0.55 mm^2, scaling this technique to RPs which have an approximate area of 1,200 mm^2, is not very feasible.

In the context of RP, cancer maps may give information on the presence and location of EPE, which is important for post surgery prognosis [40]. Several studies use RP as their training data [5, 6, 7, 41, 42, 43]. Han et al. and Gorelick et al. were papers produced by our lab [7, 43].

Han et al. has the largest data set with 340,000 mm^2 of processed tissue and Gorelick et al. had the second largest at 60,000 mm^2. The smaller datasets found in Monaco et al., DiFranco et al., Gorelick et al., Xia et al. and Khan et al. limit generalizability of their models. In this thesis, we have a data set with 340,000 mm^2 of RP tissue.

Monaco et al. use gland based classification, which may cause issues detecting high grade cancer as it commonly does not have clear glands [41]. The study uses gland lumen area and a Bayesian estimator to classify the given gland as malignant or benign. DiFranco et al. used ensemble learning to create cancer heat maps and achieved an AUC of 0.955 [42]. Gorelick et al. used an AdaBoost-based classification to label regions of interest (ROIs) [43].

Interestingly, both Xia et al. and Khan et al. use the same technique for cancer detection. They both pretrained a model using an open source breast cancer dataset from the Camelyon-16 challenge containing WSIs of lymph nodes [44], before training for prostate cancer detection [5, 6]. In both studies, WSIs were cut into 256×256 pixel ROIs which were then classified as cancer or non-cancer. Khan et al. achieves a slightly higher AUC of 0.924 and tests on a larger dataset in comparison to Xia et al. which has an AUC of 0.918. Both studies found that the pretrained model had a higher AUC than a model trained from scratch. Khan et al. found that pretrained model on the Camelyon-16 dataset performed better than a model pretrained on the ImageNet. Xia et al. uses GoogLeNet (also known as InceptionV1) [45], while Khan et al. uses InceptionV3 [46] which was built on the GoogLeNet model. These papers only test one type of model.

As shown in Table 1.1, there are several studies involving cancer detection in prostate cancer tissue. Han et al. has the largest number of patients among the RP studies and reports an AUC of 0.924 using AlexNet (a DNN published in 2012) [7, 47]. Since then numerous DNNs were created and some have higher accuracy than AlexNet when tested on the ImageNet validation dataset [34]. Without other DNNs, it is difficult to draw conclusions comparing model performance of traditional methods and DNNs. So, there is an unmet need to test multiple DNNs.

1.4 Thesis Outline

Predicting post RP outcome is important in deciding post surgery treatment. These predictions may influence the decision between adjuvant or salvage radiotherapy. Stamey et al. found tumour volume to be a significant predictor in biochemical failure [48], but Epstein et al. found that tumour volume is correlated with biochemical failure, but does not offer additional value when Gleason score and other pathologic factors were given [49]. Since then, no consensus has been reached on the significance of tumour volume on biochemical recurrence. Further research in this field requires calculating tumour volumes, which can be calculated using automated cancer maps.

The location of EPE and positive margins have been found to be important details to include in a pathology report [40, 50], therefore there is a need for automated cancer contours. These cancer maps may provide important information for post-operative targeted radiotherapy [4].

Several studies use DNNs to classify RP histology as cancer or non-cancer, but each paper only tests one type of DNN [5, 6, 7], thus there is an unmet need to training and comparing

Year **1st Author**	**Results**	**Validation Method**	**Dataset**	**Total Processed Tissue** (mm^2)	**Tissue Form**
2010 Monaco [41]	0.87 sensitivity, 0.90 specificity, pixel level	3-fold CV	40 slides 20 patients	~ 48,000	RP
2011 DiFranco [42]	0.95 AUC, ROI level	2-fold CV	15 slides 14 patients	~ 18,000	RP
2012 Doyle [36]	0.84 AUC, pixel level	3-fold CV	100 slides 58 patients	~ 1,000	Biopsy
2013 Gorelick [43]	90% Acc., ROI level	LOPO	50 slides 15 patients	~ 60,000	RP
2015 Litiens [37]	0.96 AUC, 1.0 sensitivity, 0.4 specificity, ROI level	10-fold CV	204 slides 163 patients	~ 2,040	Biopsy
2017 Kwak [21]	0.974 AUC (95% CI:0.961-0.985), sample level	491 hold-out set	653 samples	~400	TMAs from biopsies
2018 Xia [5]	0.918 AUC, 0.843 Acc., ROI level	4 patient hold-out set	16 slides 16 patients	~19,200	RP
2019 Khan [6]	0.924 AUC, ROI level	6 patient hold-out set	28 slides 28 patients	~33,600	RP
2020 Han [7]	0.964 AUC, ROI level	LOPO	284 slides 68 patients	~340,000	RP
2020 Bulten [38]	0.990 (95%CI: 0.982-0.996), biopsy level	210 biopsies hold-out set testing	5759 slides 1243 patients	~57,590	Biopsy
2020 Ström [39]	0.986 (95%CI: 0.972-0.996) AUC, biopsy level	1631 independent test set, 330 external validation set	6682 slides 976 patients	~66,820	Biopsy

Table 1.1: Cancer classification studies. Acc. indicates accuracy and LOPO indicates leave one patient out. Tissue processed area is estimated assuming 1,200 mm^2 per RP slide, 10 mm^2 per biopsy and 0.55 mm^2 per TMA sample.

multiple DNNs in cancer/non-cancer RP histology classification.

To create cancer maps we automatically generated TCMs using U-Net and tested performance across multiple centres. We then used four DNNs architectures as classification algorithms on both raw images and TCMs. By using the same training data for all models, we are able to compare model performances. We will address the questions: (1) can a DNN trained to segment slides into TCMs at one centre, accurately segment slides at another centre and (2) how do the selected DNNs compare in cancer map production and cancer detection?

Bibliography

[1] Canadian Cancer Statistics Advisory Committee, *Canadian Cancer Statistics 2019* (2019).

[2] Howlader N, Noone AM, Krapcho M, *et al.*, "SEER Cancer Statistics Review, 1975-2016.," (2019).

[3] S. Loeb, M. A. Bjurlin, J. Nicholson, *et al.*, "Overdiagnosis and overtreatment of prostate cancer," *European Urology* **65**(6), 1046–1055 (2014).

[4] J. Croke, J. Maclean, B. Nyiri, *et al.*, "Proposal of a post-prostatectomy clinical target volume based on pre-operative MRI: volumetric and dosimetric comparison to the RTOG guidelines.," *Radiation Oncology* **9**, 303 (2014).

[5] T. Xia, A. Kumar, D. Feng, *et al.*, "Patch-level Tumor Classification in Digital Histopathology Images with Domain Adapted Deep Learning," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, **2018**, 644–647 (2018).

[6] U. A. H. Khan, C. Stürenberg, O. Gencoglu, *et al.*, "Improving Prostate Cancer Detection with Breast Histopathology Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11435**, 91–99 (2019).

[7] W. Han, C. Johnson, M. Gaed, *et al.*, "Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens," *Scientific Reports* **10**(1), 9911 (2020).

[8] M. G. Sanda, J. A. Cadeddu, E. Kirkby, *et al.*, "Clinically Localized Prostate Cancer: AUA/ASTRO/SUO Guideline. Part II: Recommended Approaches and Details of Specific Care Options," *Journal of Urology* **199**(4), 990–997 (2018).

[9] M. Han, A. W. Partin, M. Zahurak, *et al.*, "Biochemical (prostate specific antigen) recurrence probability following radical prostatectomy for clinically localized prostate cancer," *Journal of Urology* **169**(2), 517–523 (2003).

[10] T. M. Pisansky, I. M. Thompson, R. K. Valicenti, *et al.*, "Adjuvant and Salvage Radiotherapy after Prostatectomy: ASTRO/AUA Guideline Amendment 2018-2019," *The Journal of Urology* **202**(3), 533–538 (2019).

[11] W. J. Catalona, G. F. Carvalhal, D. E. Mager, *et al.*, "Potency, continence and complication rates in 1,870 consecutive radical retropubic prostatectomies," *Journal of Urology* **162**(2), 433–438 (1999).

[12] I. M. Thompson, R. K. Valicenti, P. Albertsen, *et al.*, "Adjuvant and salvage radiotherapy after prostatectomy: AUA/ASTRO guideline," *Journal of Urology* **190**(2), 441–449 (2013).

[13] S. C. Morgan, T. S. Waldron, L. Eapen, *et al.*, "Adjuvant radiotherapy following radical prostatectomy for pathologic T3 or margin-positive prostate cancer: A systematic review and meta-analysis," **88**(1), 1–9 (2008).

[14] T. Wiegel, D. Bartkowiak, D. Bottke, *et al.*, "Adjuvant radiotherapy versus wait-and-see after radical prostatectomy: 10-year follow-up of the ARO 96-02/AUO AP 09/95 trial," *European Urology* **66**, 243–250 (2014).

[15] K. S. Suvarna, C. Layton, and B. J. D., *Bancroft's Theory and Practice of Histological Techinques*, Elsevier (2013).

[16] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology* **23**(4), 291–299 (2001).

[17] W. Han, A. D. Ward, C. Johnson, *et al.*, "Automatic cancer detection and localization on prostatectomy histopathology images ," in *Medical Imaging 2018: Digital Pathology*, J. E. Tomaszewski and M. N. Gurcan, Eds., **10581**, 205–212, International Society for Optics and Photonics, SPIE (2018).

[18] A. M. Khan, N. Rajpoot, D. Treanor, *et al.*, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering* **61**(1), 1729–1738 (2014).

[19] J. Vicory, H. D. Couture, N. E. Thomas, *et al.*, "Appearance normalization of histology slides," *Computerized Medical Imaging and Graphics* **43**(1), 89–98 (2015).

[20] M. Macenko, M. Niethammer, J. S. Marron, *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, (2009).

[21] J. T. Kwak and S. M. Hewitt, "Nuclear Architecture Analysis of Prostate Cancer via Convolutional Neural Networks," *IEEE Access* **5**, 18526–18533 (2017).

[22] N. Chen and Q. Zhou, "The evolving gleason grading system," *Chinese Journal of Cancer Research* **28**(4), 58–64 (2016).

[23] J. I. Epstein, W. C. Allsbrook, M. B. Amin, *et al.*, "The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma," *American Journal of Surgical Pathology* **29**(9), 1228–42 (2005).

[24] J. I. Epstein, L. Egevad, M. B. Amin, *et al.*, “The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma,” *American Journal of Surgical Pathology* **40**, 244–252 (2016).

[25] W. C. Allsbrook, K. A. Mangold, M. H. Johnson, *et al.*, “Interobserver reproducibility of Gleason grading of prostatic carcinoma: Urologic pathologists,” *Human Pathology* **32**(1), 74–80 (2001).

[26] S. Ekici, A. Ayhan, I. Erkan, *et al.*, “The Role of the Pathologist in the Evaluation of Radical Prostatectomy Specimens,” *Scandinavian Journal of Urology and Nephrology* **37**(5), 387–391 (2003).

[27] L. Egevad, A. S. Ahmad, F. Algaba, *et al.*, “Standardization of Gleason grading among 337 European pathologists,” *Histopathology* **62**(2), 247–256 (2013).

[28] J. K. McKenney, J. Simko, M. Bonham, *et al.*, “The potential impact of reproducibility of gleason grading in men with early stage prostate cancer managed by active surveillance: A multi-institutional study,” *Journal of Urology* **186**(2), 465–469 (2011).

[29] T. A. Ozkan, A. T. Eruyar, O. O. Cebeci, *et al.*, “Interobserver variability in Gleason histological grading of prostate cancer,” *Scandinavian Journal of Urology* **50**(6), 420–424 (2016).

[30] T. H. Van Der Kwast, L. Collette, H. Van Poppel, *et al.*, “Impact of pathology review of stage and margin status of radical prostatectomy specimens (EORTC trial 22911),” *Virchows Archiv* **449**(4), 428–434 (2006).

[31] A. J. Evans, P. C. Henry, T. H. Van Der Kwast, *et al.*, “Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens,” *American Journal of Surgical Pathology* **32**(10), 1503–1512 (2008).

[32] W. Bulten, M. Balkenhol, J. J. A. Belinga, *et al.*, “Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists,” *Modern Pathology* (2020).

[33] A. G. Ivakhnenko and V. G. Lapa, *Cybernetics and forecasting techniques*, Modern analytic and computational methods in science and mathematics, American Elsevier Pub. Co. (1967).

[34] J. Deng, W. Dong, R. Socher, *et al.*, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).

[35] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, **9351**, 234–241 (2015).

[36] S. Doyle, M. Feldman, J. Tomaszewski, *et al.*, “A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies,” *IEEE Transactions on Biomedical Engineering* **59**(5), 1205–18 (2012).

[37] G. Litjens, B. E. Bejnordi, N. Timofeeva, *et al.*, “Automated detection of prostate cancer in digitized whole-slide images of H and E-stained biopsy specimens,” **9420**, 64–69 (2015).

[38] W. Bulten, H. Pinckaers, H. van Boven, *et al.*, “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study,” *The Lancet Oncology* **21**(2), 233–241 (2020).

[39] P. Ström, K. Kartasalo, H. Olsson, *et al.*, “Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study,” *The Lancet Oncology* **21**(2), 222–232 (2020).

[40] C. Magi-Galluzzi, A. J. Evans, B. Delahunt, *et al.*, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 3: Extraprostatic extension, lymphovascular invasion and locally advanced disease," *Modern Pathology* **24**(1), 26–38 (2011).

[41] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, *et al.*, "High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models," *Medical Image Analysis* **14**(4), 617–29 (2010).

[42] M. D. DiFranco, G. O'Hurley, E. W. Kay, *et al.*, "Ensemble based system for whole-slide prostate cancer probability mapping using color texture features," *Computerized Medical Imaging and Graphics* **35**(7-8), 629–645 (2011).

[43] L. Gorelick, O. Veksler, M. Gaed, *et al.*, "Prostate histopathology: Learning tissue component histograms for cancer detection and classification," *IEEE Transactions on Medical Imaging* **32**(10), 1804–1818 (2013).

[44] "ISBI challenge on cancer metastasis detection in lymph node - Camelyon16," (2016).

[45] C. Szegedy, Wei Liu, Yangqing Jia, *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9 (2015).

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826 (2016).

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, *et al.*, Eds., 1097–1105, Curran Associates, Inc. (2012).

[48] T. A. Stamey, J. E. McNeal, C. M. Yemoto, *et al.*, “Biological determinants of cancer progression in men with prostate cancer,” *Journal of the American Medical Association* **281**(15), 1395–1400 (1999).

[49] J. I. Epstein, M. Carmichael, A. W. Partin, *et al.*, “Is tumor volume an independent predictor of progression following radical prostatectomy? A multivariate analysis of 185 clinical stage B adenocarcinomas of the prostate with 5 years of followup,” *Journal of Urology* **149**(6), 1478–1481 (1993).

[50] P. H. Tan, L. Cheng, J. R. Srigley, *et al.*, “International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 5: Surgical margins,” *Modern Pathology* **24**(1), 48–57 (2011).

Chapter 2

Cancer Detection on Digitized Prostatectomy Slides using Convolutional Neural Networks

Abstract

Purpose: Deep learning algorithms have been proposed as a tool to assist the pathologist with slide annotation. Before they can be used, several models should be compared to identify which, if any, model performs better than others. **Approach**: We used a U-Net to automatically segment tissue components (nuclei, lumina and stroma/other components) from digitized radical prostatectomy haematoxylin and eosin stained whole-mount tissue samples from 68 patients, using 2-fold and 5-fold cross validation. We trained AlexNet, NASNet, VGG16 and Xception, all pretrained on the ImageNet database, to classify regions of interest (ROIs) as cancer or non-cancer using either tissue component maps (TCMs) generated by the U-Net, or raw images. We measured classifier performance using leave-one-patient-out cross validation. **Results**: The U-Net segmented lumen and nuclei with an area under the receiver operating characteristic curve (AUC) of 0.975 (95% confidence interval 0.971-0.981) and 0.983 (0.981-0.985) respectively. On an open-source dataset, the U-Net achieved an AUC of 0.985 (0.982-0.989) and 0.999 (0.999-0.999) for lumen and nuclei respectively. For cancer detection, VGG16 trained on raw images both performed with the highest AUC of 0.994 (0.992-0.996) and the lowest false negative rate of 0.048 (0.027-0.065). A voting scheme consisting of the four models trained on raw images performed with the lowest false positive rate of 0.017 (0.012-0.020) and error rate of 0.020 (0.013-0.026). **Conclusions**: The U-Net can be trained on images from one centre and then accurately segment images from another centre into TCMs. Multiple deep neural networks can be trained to accurately classify ROIs as cancer vs. non-cancer with minor differences in error metrics. A voting scheme consisting of multiple deep networks reduces the false positive rate.

2.1 Introduction

Prostate cancer (PCa) is the third most frequently-diagnosed cancer globally [1]. One of the treatment options is radical prostatectomy (RP), which is typically performed for those who have organ-confined cancer. Post surgery information from tissue pathology is important in predicting prognosis [2] through the location of extraprostatic extension (EPE) and positive margins [3, 4], and has potential to guide post-operative targeted radiotherapy [5]. This means pathology reports may benefit from including cancer maps created from histology slides.

Approximately 35% of patients experience biochemical recurrence after RP [6], thus it is important to predict who would benefit from post surgery treatments. Tumour volume may be a predictor of post surgery outcome but, a consensus has not been reached [7, 8]. Unfortunately, tumour volume estimates using detailed contours are labour intensive which limits feasibility of studies involving tumour volumes.

One treatment option is adjuvant radiotherapy, where radiotherapy is done immediately post surgery. Another option is salvage radiotherapy, where the patient undergoes active surveillance and radiotherapy is done if recurrence is detected. It is important to determine which patients will most benefit from adjuvant radiotherapy, since radiotherapy can cause complications like gastrointestinal and genitourinary toxicity, urinary incontinence, and a debated increase in risk for secondary malignancies [9, 10, 11]. Debate as to which treatment is most beneficial is still ongoing. The American Urological Association (AUA) and the American Society for Radiation Oncology (ASTRO) concluded after reviewing studies on adjuvant and salvage radiotherapy, that there was not sufficient evidence to determine which treatment is superior [9, 12].

Deep learning is a popular technique for computer vision problems and was used to classify images as cancer or non-cancer in RP whole slide images (WSIs) [13, 14, 15]. Xia et al. and Khan et al. used a smaller patient set than what is used in this paper. Han et al. used the same dataset used in this paper, but only tests one deep neural network (DNN) called AlexNet [16]. AlexNet was published in 2012 and since then many DNN architectures were published. To the best of our knowledge, there has not been a study that compares multiple DNNs trained to classify cancer on RP WSIs.

Due to the layers of complexity found in deep networks, it is often difficult to interpret the results. When we limit the amount of input information provided to the networks by using tissue component maps (TCMs) and compare to performance of a model trained on raw images, we can gain understanding as to the primary aspects of the tissue needed by the neural network to determine if something is cancer or non-cancer. TCMs are simply the raw histology images segmented into different tissues. Several studies have shown that nuclei maps or other tissue component maps result in similar cancer detection accuracy to raw image inputs in a deep learning context [14, 17].

In this work, we applied deep learning to: (1) segment tissue components from histology slides and (2) compare error metrics of four DNN models trained to label regions of interest (ROIs) as cancer or non-cancer.

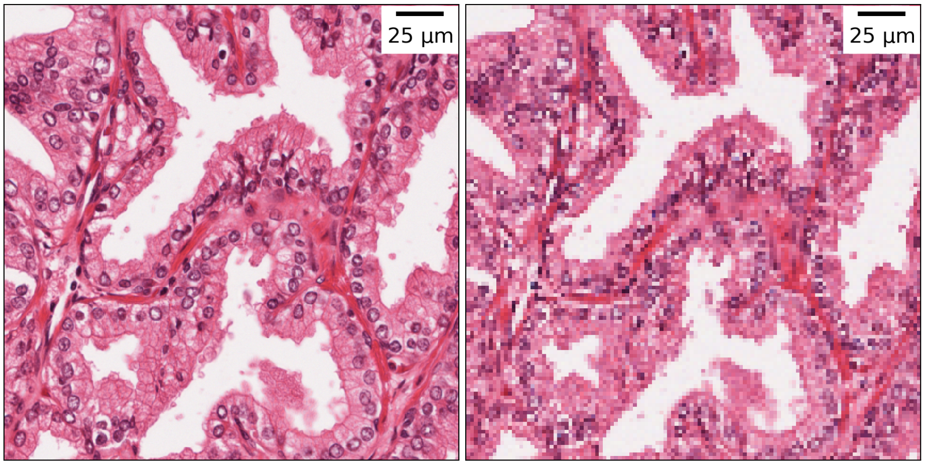
2.2 Methods

2.2.1 Materials

From 68 patients enrolled in our Image Guidance for Prostate Cancer (IGPC) trial, we obtained 286 WSIs of whole-mount, mid-gland radical prostatectomy tissue sections, stained with haematoxylin and eosin (H&E). The slides were scanned at 0.5 μm/pixel resolution using an Aperio ScanScope at 20 × magnification. We then downsampled to 2 μm/pixel using nearest neighbour approximation and digitally cut into 240×240 pixel (480 μm×480 μm) ROIs, resulting in approximately 1.3 million ROIs. An example of the full resolution and downsampled image can be found on Figure 2.1. These whole-mount images were contoured and graded at full resolution by a trained physician—taking approximately 70 hours per patient. The contours were then verified by one of two pathologists. Each ROI was classified as cancer if more than 50% of the area was cancer, according to the ground-truth contours.

We also obtained open source slides from The Cancer Imaging Archive (TCIA) [18, 19], which included 16 patients with a total of 114 digitally restitched psuedo-whole mount slides. Of the slides, we excluded 4 as they were visually at a significantly lower resolution than the ROIs used for training. The spatial resolution of these images was not provided and the cancer was not contoured or graded. Therefore we used these 110 slides as an external dataset only to validate the TCMs produced by the U-Net.

Our laboratory previously developed an adaptive thresholding based method to generate TCMs by categorizing each pixel in the image as belonging either to a nucleus, lumen, or other tissue region [20]. Although this method has demonstrated robustness to staining variability, it is currently limited to the aforementioned three-way classification and depends on a per-slide calibration step that requires computational time. In principle, a deep neural network (DNN) trained to perform tissue component mapping could eliminate the calibration step and could be readily extended to find a wider array of different tissue components, potentially supporting more accurate cancer detection and grading. As a first step, we sought to train a DNN to create



(a) Full Resolution (b) Down sampled

Figure 2.1: Full resolution and down sampled images using nearest neighbour approximation where the resolution of (a) is 0.5 μm/px and (b) is 2 μm/px. Both images correspond to 480 μm by 480 μm area of tissue.

three-way TCMs, using TCMs generated from an adaptive thresholding approach as training data. This approach has the advantage of generating a much larger number of training TCMs than what would be possible to produce by manual annotation.

To create ground truth TCMs, we arbitrarily selected two ROIs per patient for the IGPC validation dataset, and two ROI per slide for the TCIA dataset. This means there was a total of 136 ROIs used for validation from the IGPC dataset and 220 for the TCIA dataset. We validated the segmentations against a subset of TCM ROIs generated by manual thresholding the blue and green channels to segment the nuclei and luminal regions, respectively. These maps were then manually updated using a paintbrush tool to remove any false lumina, or RBCs falsely classified as nuclei. This was done by a graduate student.

Models were written using the Keras v2.2.4 library in Python 3.6.8, or MATLAB 2017a. The network was trained on the GeForce GTX 1080 Ti, with NVIDIA Driver v419.67 and CUDA Toolkit v10.0.130. Statistical analysis was done on Python using Scipy v1.4.1.

2.2.2 Network Architecture

We trained a U-Net [21] to produce TCMs from unprocessed, H&E stained, histology images from the IGPC dataset. Each U-Net was trained to create either nuclei or lumen maps. We then combined the outputs of the trained networks to create three-tissue TCMs, containing nuclei, lumina, or stromal/other tissue regions. If a pixel in the map was labelled as both nucleus and lumen, then the pixel was labelled nuclei because the false positive rate (FPR) of nuclei was lower than for lumen. This means a positive nucleus detection is more likely than a positive lumen detection.

We used a weighted binary cross entropy as the loss function for the DNN trained to produce nuclei maps. In the original U-Net paper, a custom weighted loss function was used to encourage separation of adjacent objects and to account for class imbalance. For our project, nucleus separation was not an issue so we only applied a weighting factor to address the class imbalance for nucleus pixels. The proportion of lumen to background was relatively balanced, thus we did not add any weights to the binary cross entropy loss.

The weighted loss is defined in eq. 2.1, where y is the expected label and $p(y)$ is the probability of getting the expected value. The weighting factors (W_+ and W_-) were calculated using the number of positive pixels (N_+), negative pixels (N_-) and the total number of pixels (N) found in each batch, where $W_+ = \frac{N_+}{N}$ and $W_- = \frac{N_-}{N}$. When applied to the loss function we get

$$L_w(y) = -\frac{1}{N}\sum_{i=0}^{N} W_+ \cdot y_i \cdot log(p(y_i)) + W_- \cdot (1 - y_i) \cdot log(1 - p(y_i)) \tag{2.1}$$

where W_- is the weight of a negative category and W_+ is the weight of a positive category. Every misclassification of a nuclei pixel as a background pixel has a high cost to the loss function, so the model is deterred from labelling the entire image as background.

2.2.3 Training

We trained the U-Net using a 2-fold and 5-fold cross validation scheme to randomly separate a set of patients for testing. The remaining set of patients were then randomly split into training and validation, where 70% were put into the training and 30% to the validation set. We ensured each patient only appears in one set. We chose a batch size of 16, an adaptive moment estimator optimizer called Adam [22] and a learning rate of 1e-5. These hyperparameters were selected to optimize for shorter training time and higher accuracy on the validation data.

To classify ROIs as cancer or non-cancer, we decided to use a range of pretrained networks which use different model architecture. AlexNet [16], NASNet [23], VGG16 [24] and Xception [25] were selected. These networks use different deep learning techniques where AlexNet is one of the first DNNs that use convolutional kernels, VGG16 improves on the AlexNet model by using smaller kernels, Xception uses modified depthwise separable convolutions and NASNet uses neural architecture search. Since AlexNet was not available in Keras, MATLAB was used instead. All four models were pretrained on the ImageNet dataset, an open-source project containing 14 million images intended for image and vision research [26]. For the training data, we randomly removed non-cancer ROIs, to ensure the number of cancer ROIs is equal to non-cancer for each patient.

We trained VGG16 and Xception using 240×240×3 pixel sized ROIs, while the images used to train AlexNet and NASNet were resized using bilinear interpolation to 227×227×3 pixels and 224×224×3 pixels respectively. AlexNet had a fixed input size due to limitations in MATLAB's deep learning implementation and NASNet was fixed due to architecture design.

We validated the cancer detection models using a leave-one-patient-out (LOPO) cross validation scheme, where a patient is left out for testing. The remaining patients were then randomly split into 70% training and 30% validation. Hyperparameters are summarized in Table 2.1 and were set using the validation dataset. For AlexNet, the learning rate was 1e-4 for all layers except the last layer which had a learning rate of 2e-3. All networks used Adam optimizer [22].

Model	**Loss Function**	**Optimizer**	**Batch Size**	**Epochs**	**Input Size (pixels)**
AlexNet-TCM	BC	Adam, lr = 1e-4	200	10	227×227×3
AlexNet-RAW	BC	Adam, lr = 1e-4	200	10	227×227×3
NASNet-TCM	BC	Adam, lr = 1e-4	128	10	224×224×3
NASNet-RAW	BC	Adam, lr = 1e-4	32	10	224×224×3
VGG16-TCM	BC	Adam, lr = 1e-5	32	10	240×240×3
VGG16-RAW	BC	Adam, lr = 1e-5	32	10	240×240×3
Xception-TCM	BC	Adam, lr = 1e-4	32	10	240×240×3
Xception-RAW	BC	Adam, lr = 1e-4	32	10	240×240×3

Table 2.1: Model hyperparameters. Lr indicates learning rate and BC binary crossentropy. Adam refers to the adaptive moment optimizer [22].

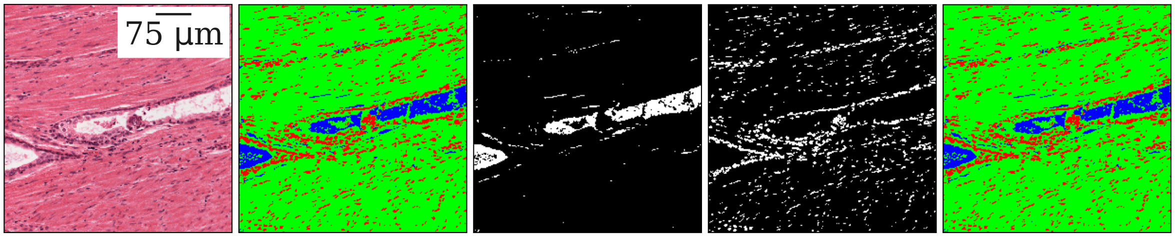
To compare TCMs against raw images, we trained the four networks (AlexNet, NASNet, VGG16 and Xception) on raw images or TCMs generated from the 2-fold cross validation scheme, resulting in a total of 8 trained networks. The same training, validation and testing split was used across all networks to remove data distribution as a confounding factor when comparing the models. These models will be referred to as AlexNet-TCM, NASNet-TCM, VGG16-TCM and Xception-TCM for those trained on TCMs and AlexNet-RAW, NASNet-RAW, VGG16-RAW and Xception-RAW for those trained on raw images.

These models were combined in a majority voting ensemble model. Predictions were given on a majority vote (>50%) from all models trained on raw images. The same was done for TCM models and for a combination of all eight models. We refer to them as voting-RAW, voting-TCM and voting-All for majority voting with only models trained on raw images, only on TCMs and all models respectively.

2.3 Results

2.3.1 Tissue Component Maps

Training a U-Net required approximately 12 hours of training time for lumen and 24 hours for nuclei with GPU acceleration. Producing a TCM for an entire slide after the model is trained takes approximately 3 minutes with a GPU. An example of a final tissue component map can be found at Fig.2.2.



(a) Histology (b) Training TCM (c) U-Net lumen (d) U-Net nuclei (e) U-Net TCM

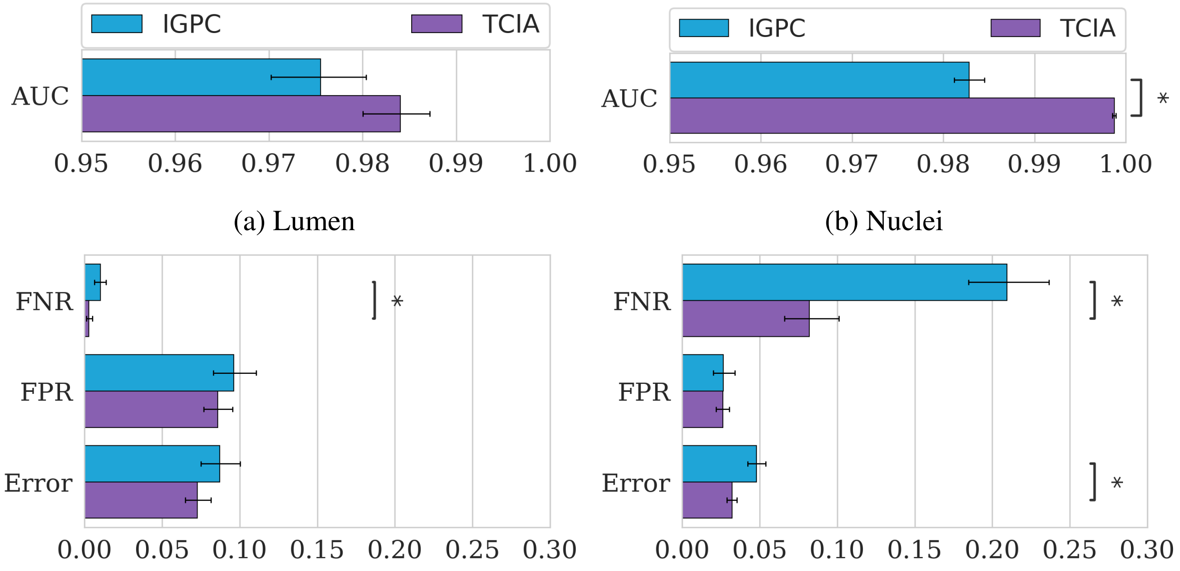
Figure 2.2: TCM generated using U-Nets (e), compared against TCMs generated using adaptive thresholding (b). In (b) and (e) red indicates nuclei, blue lumen and green other. In (c) and (d) black indicates background and white indicates the corresponding tissue.

In this paper we will use FNR for false negative rate, FPR for false positive rate and AUC for area under the receiver operating characteristic curve. We will be reporting the mean and 95% confidence interval (CI) of each performance metric.

For the IGPC dataset, the AUC was 0.970 (0.965-0.976) for lumen and 0.967 (0.959-0.975) for nuclei in the 5-fold cross validation, and 0.975 (0.971-0.981) for lumen and 0.983 (0.981-0.984) for nuclei in the 2-fold validation. For the TCIA dataset, the respective AUCs were 0.980 (0.977-0.984) for lumen and 0.994 (0.993-0.996) for nuclei from a 5-fold validation, and 0.985 (0.982-0.989) for lumen and 0.999 (0.999-0.999) for nuclei from a 2-fold validation. TCIA had higher AUCs than for the IGPC dataset, despite the fact that the networks were trained on the IGPC dataset. Other error metrics can be found in Table 2.2 and in Figure 2.3. Since none of the TCIA data was used for training, we calculated error metrics per fold and averaged as an overall mean per ROI.

Model	**FNR**	**FPR**	**Error Rate**	**AUC**
5-fold cross validation				
TCIA-lumen	0.003 (0.001-0.005)	0.086 (0.076-0.095)	0.073 (0.065-0.081)	0.980 (0.977-0.984)
TCIA-nucleus	0.084 (0.068-0.100)	0.021 (0.017-0.025)	0.029 (0.026-0.032)	0.994 (0.993-0.995)
IGPC-lumen	0.010 (0.006-0.013)	0.100 (0.087-0.113)	0.091 (0.078-0.103)	0.970 (0.965-0.976)
IGPC-nucleus	0.219 (0.193-0.243)	0.023 (0.018-0.028)	0.047 (0.041-0.053)	0.967 (0.959-0.975)
2-fold cross validation				
TCIA-lumen	0.003 (0.001-0.005)	0.084 (0.075-0.093)	0.071 (0.063-0.079)	0.985 (0.982-0.989)
TCIA-nucleus	0.080 (0.063-0.096)	0.027 (0.022-0.032)	0.033 (0.029-0.037)	0.999 (0.999-0.999)
IGPC-lumen	0.010 (0.006-0.014)	0.097 (0.083-0.110)	0.088 (0.075-0.099)	0.975 (0.971-0.981)
IGPC-nucleus	0.210 (0.185-0.235)	0.024 (0.019-0.029)	0.046 (0.041-0.050)	0.983 (0.981-0.984)

Table 2.2: Segmentation mean and 95% CI of TCMs generated with a U-Net, validated against validation ROIs segmentations generated by manual thresholding. 136 ROIs from TCIA and 220 from TCIA were used for validation.



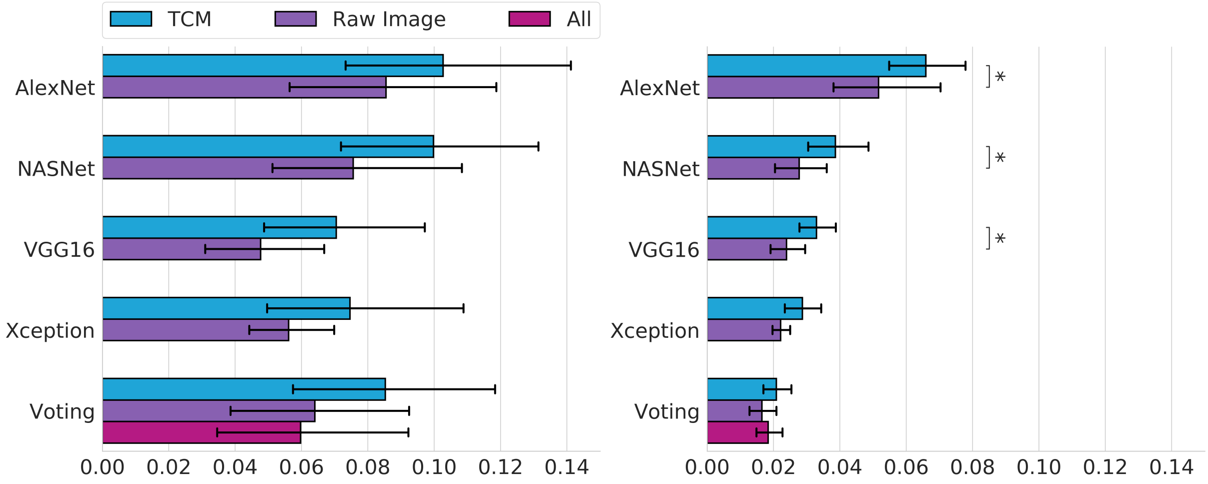
(a) Lumen (b) Nuclei

(c) Lumen (d) Nuclei

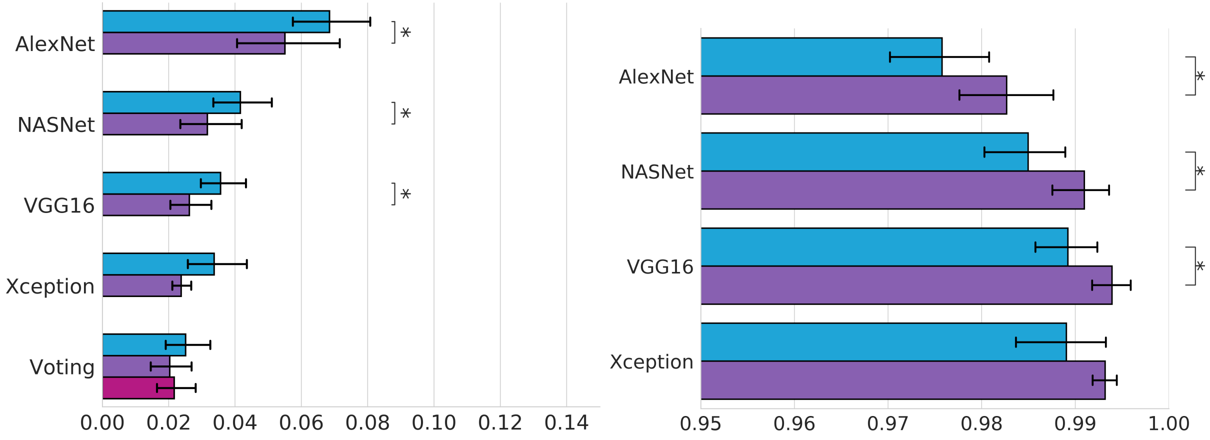
Figure 2.3: Error metrics for nuclei and lumen segmentation using 2-fold U-Net on either the TCIA or IGPC data in a boxplot where error bars represent 95% CI. “*” indicates a p-value less than 0.05 using a 2-tailed Mann-Whitney-Wilcoxon test.

2.3.2 Cancer Detection

The TCMs generated via the 2-fold cross validation U-Nets were then used for cancer vs. non-cancer classification using a LOPO cross validation scheme. Training the classifiers took approximately 2 hours per fold, for a total of 7 days per model. Labelling ROIs for one WSI post training took approximately 2-4 minutes with GPU acceleration. Error values can be found in Table 2.3 and Figure 2.4. Error metrics were also calculated using a majority voting scheme from all of the models trained on either raw images, or TCMs or out of all eight models. The ensemble models did not have AUCs, as their labels were binary. The model with the lowest AUC was AlexNet-TCM with 0.976 (0.971-0.982) and the highest AUC was VGG16-RAW with 0.994 (0.992-0.996).



(a) FNR (b) FPR



(c) Error Rate (d) AUC

Figure 2.4: Error metrics of cancer classification from a 2-fold cross validation scheme."*" indicates a p-value less than 0.05 using a 2-tailed Mann-Whitney-Wilcoxon test. Error metrics were calculated on a per-patient basis. Error bars indicate 95% CI.

Network	Input	Error Metric			
		FNR	**FPR**	**Error Rate**	**AUC**
AlexNet	TCM	0.103 (0.067-0.134)	0.066 (0.053-0.077)	0.068 (0.056-0.079)	0.976 (0.971-0.982)
	RAW	0.085 (0.050-0.115)	0.052 (0.034-0.066)	0.055 (0.038-0.070)	0.983 (0.978-0.988)
NASNet	TCM	0.100 (0.068-0.128)	0.039 (0.029-0.047)	0.042 (0.032-0.050)	0.985 (0.981-0.990)
	RAW	0.076 (0.045-0.101)	0.028 (0.020-0.036)	0.032 (0.022-0.040)	0.991 (0.988-0.994)
VGG16	TCM	0.070 (0.045-0.096)	0.033 (0.027-0.039)	0.036 (0.029-0.042)	0.989 (0.986-0.993)
	RAW	**0.048 (0.027-0.065)**	0.024 (0.018-0.029)	0.026 (0.020-0.032)	**0.994 (0.992-0.996)**
Xception	TCM	0.075 (0.040-0.103)	0.029 (0.023-0.034)	0.034 (0.023-0.042)	0.989 (0.985-0.995)
	RAW	0.056 (0.043-0.068)	0.022 (0.019-0.025)	0.024 (0.021-0.027)	0.993 (0.992-0.994)
Voting	TCM	0.085(0.053-0.113)	0.021(0.016-0.025)	0.025(0.018-0.031)	—
	RAW	0.064(0.035-0.088)	**0.017(0.012-0.020)**	**0.020(0.013-0.026)**	—
	All	0.060(0.030-0.084)	0.018(0.014-0.022)	0.022(0.015-0.027)	—

Table 2.3: Cancer detection error metrics mean and 95% CI which were calculated on a per patient basis. Bold-face indicates the measurement with the lowest mean of the columns FNR, FPR and Error Rate and the highest for AUC.

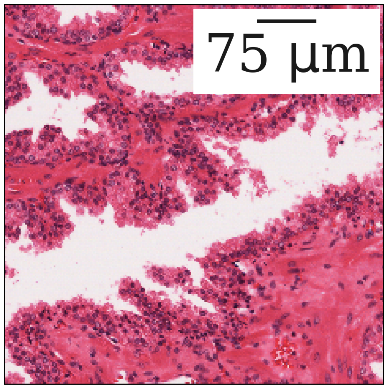
2.4 Discussion

2.4.1 U-Net Segmentations

Examination of sample ROIs yields a better understanding of how the U-Net performs. The U-Net tended to over-segment lumen, as is evident from the high number of false positives in Figure 2.3 and Figure 2.6. This may be an artifact from the training data, since the training data also over-segmented lumen as seen in Figure 2.6. In this example, there is a large amount of false positives in both U-Net and adaptive thresholding based segmentations. The adaptive thresholding technique used a dynamically selected threshold for nuclei, but a fixed global threshold to separate lumen from all other tissue. Thresholding is unable to differentiate slide background caused by tissue tears and slide background from lumina. Since this example is lightly stained, portions where the tissue is lighter in colour are classified as lumina.

It is possible that the U-Net will be able to differentiate tissue tears from lumen if given properly segmented training data. Unlike thresholding techniques, the U-Net may incorporate the distribution and shape of nuclei and other tissue, to identify lumina.

By comparing Figure 2.5 and Figure 2.6, we can see differences between model performance on examples with different staining intensities. The U-Net performs reasonably well



(a) histology

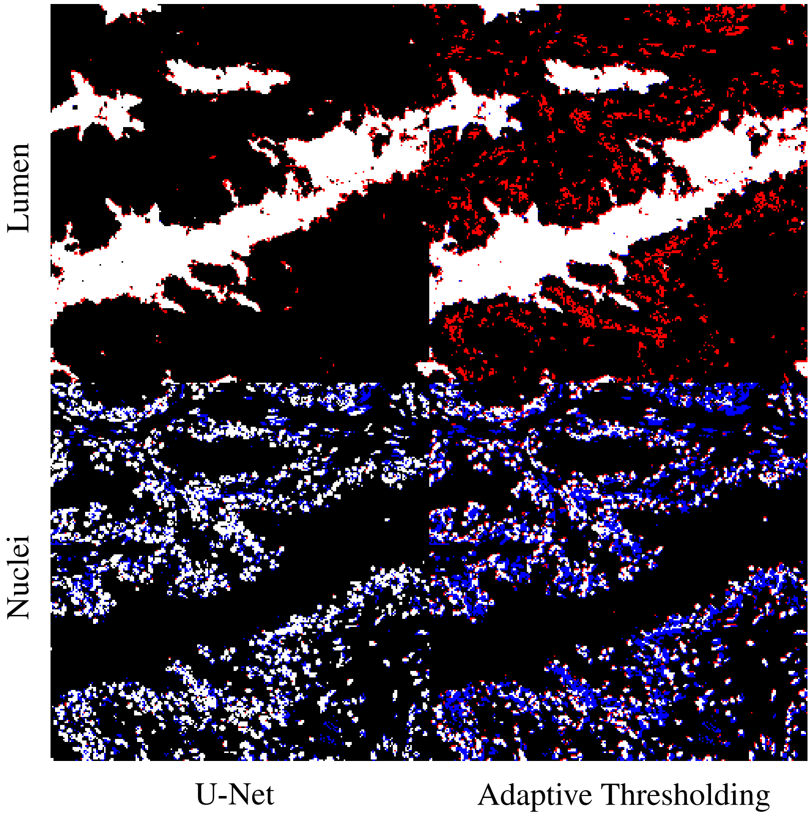
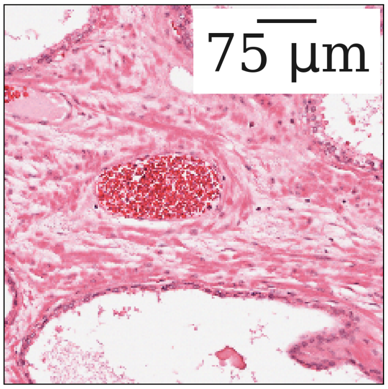


Figure 2.5: Error maps of nuclei and lumen generated from U-Net and adaptive thresholding where white is true positive, red is false positive, black is true negative and blue is false negative. The segmentations were compared against manual thresholding.



(a) histology

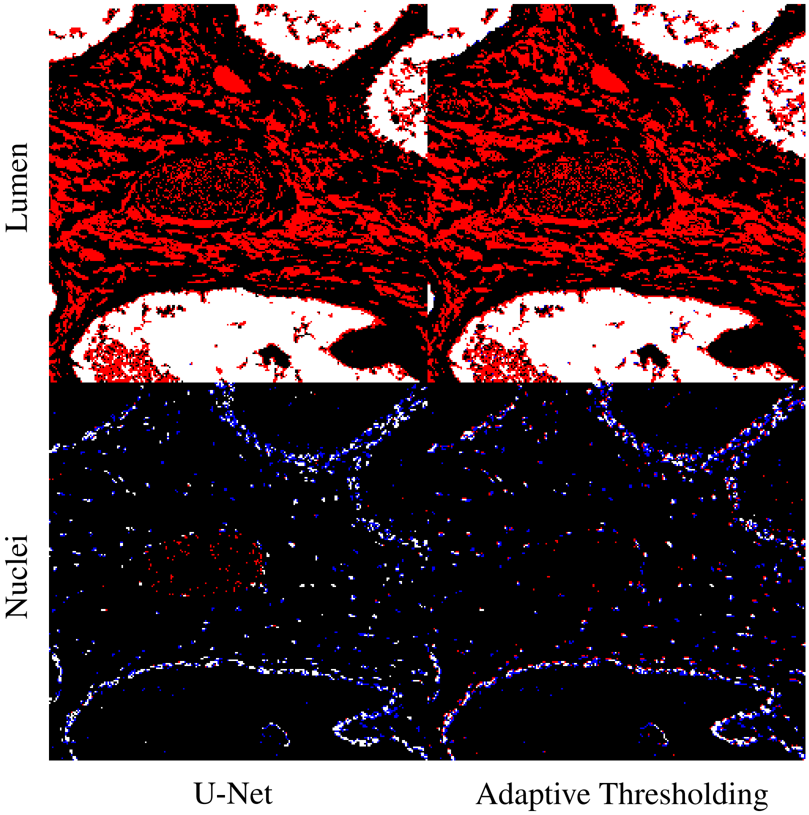


Figure 2.6: Error maps of nuclei and lumen generated from U-Net and adaptive thresholding where white is true positive, red is false positive, black is true negative and blue is false negative, for a lightly stained example. The segmentations were compared against manual thresholding.

for both examples, with the darkly stained example in Figure 2.5, performing with fewer false negative nuclei and false positive lumen than the lightly stained example.

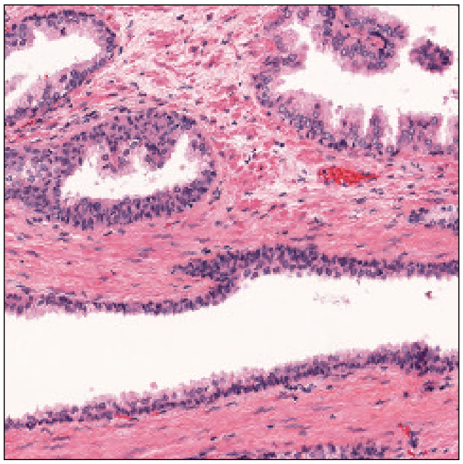
Figure 2.6 has more red blood cells (RBCs) that were falsely classified as nuclei in the U-Net segmentation, compared to adaptive thresholding. The adaptive threshold technique removed RBCs explicitly using hue-saturation-intensity thresholds. The U-Net also only misclassified a small portion of the RBCs in this example and does not require this second thresholding step. It is possible that if the U-Net was trained to mask RBCs, its performance would be comparable to adaptive thresholding.

The U-Net nuclei segmentation in Figure 2.5 had fewer false negatives than the adaptive thresholding technique and fewer false positive lumen. The U-Net was trained to imitate the adaptive thresholding technique, so we expected it to output segmentations nearly identical to adaptive thresholding if given enough epochs. Since this model was only trained for 10 epochs, it is possible that if it was trained for longer, the model would perform nearly identically to adaptive thresholding. Instead, U-Net may have reached a compromise that balances between the different staining intensities found in the training data, resulting in a more accurate segmentation.

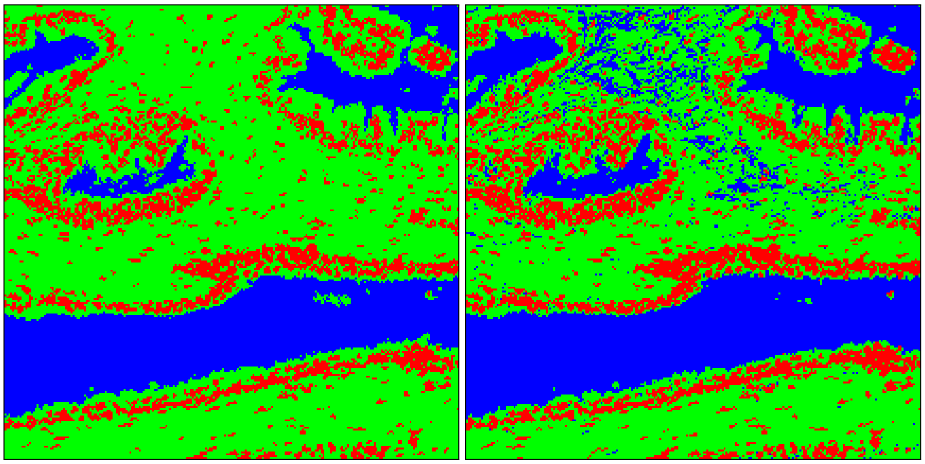
Validation on an external dataset showed that U-Net can provide equivalent performance on images acquired at a different centre to that in which it was trained. Figure 2.7 is an example ROI taken from the external dataset and its resulting output segmentations. Qualitatively, this example performed similarly to the IGPC database. Only a portion of the RBCs were not falsely classified as nuclei. Like Figure 2.6 which has a similar staining intensity, the U-Net over segmented lumen.

Looking at Figure 2.3 we see that in general, error metrics from TCIA data are similar to IGPC. There was no statistically significant difference between TCIA and IGPC in lumen AUC, FPR and error rate, and nuclei FPR. TCIA data had a higher nuclei AUC, lower nuclei FPR, error rate and lower lumen FNR. Since the IGPC dataset used whole-mount tissue samples, the H&E staining was done manually rather than robotically, resulting in large staining variability.

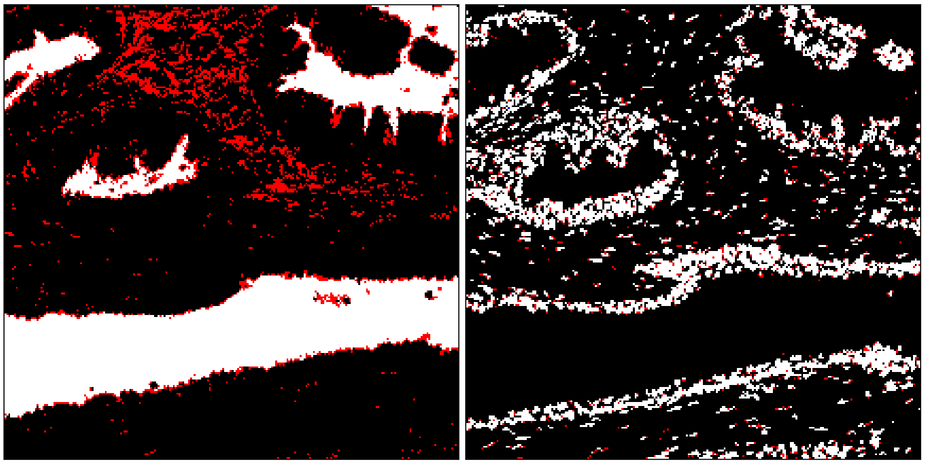
This may account for the larger variation in error metrics and poorer performance in comparison to TCIA. Although information on the exact staining procedure was not provided for the TCIA dataset, the tissue was cut into quadrants and the qualitative uniformity of the samples imply robotic staining, as is normally conducted in clinical pathology labs for standard-sized histology slides.



(a) histology



(b) Manual TCM (c) U-Net TCM



(d) U-Net lumen error map (e) U-Net nuclei error map

Figure 2.7: TCM generated using U-Net, compared against TCMs generated using manual thresholding for TCIA data. In (b) and (c) red indicates nuclei, blue lumen and green other in the TCMs. In (d) and (e) white is true positive, red is false positive, black is true negative and blue is false negative. The resolution of these images was not reported in the TCIA database.

2.4.2 Cancer Detection

The error metrics for cancer detection have been plotted in Figure 2.4. Networks trained on raw images performed similarly to those trained on TCMs, but raw image models consistently performed with a lower FNR, FPR, error rate and a higher AUC.

AlexNet performed with the highest FNR, FPR, error rate and the lowest AUC of the four networks, although not by a large margin. It is also the oldest network. Advancements in network architecture may have resulted in the poorer performance of AlexNet in comparison to newer networks. VGG16 uses a very similar structure to AlexNet but uses a smaller kernel size. Xception relies on depth-wise separable convolutional layers, which uses convolution in a method that reduces the number of parameters. NASNet optimizes network architecture as a part of its training process. These changes may have improved performance. It is also possible that the difference between AlexNet and the other models is caused by the differences in MATLAB's deep learning toolbox and Keras. Although both libraries use the ImageNet dataset to pretrain the models, details in pretraining methods like batch size and number of epochs may affect the resulting models.

VGG16-RAW performed with highest AUC of 0.994(0.992-0.996), while Xception-RAW had the next highest AUC of 0.993 (0.992-0.994). The difference between AUCs is minimal and not significant enough to show a clear preference for one model. Out of the eight DNNs, VGG16-RAW had the lowest FPR at 0.048 (0.027-0.065).

When we include the majority voting schemes, voting-RAW had the lowest FPR at 0.017 (0.012-0.020) and error rate at 0.020 (0.013-0.026). The voting scheme likely reduces false positives, as multiple networks have to agree for voting to output a positive. When we look at Figure 2.9 and Figure 2.10, we see that cancer maps produced by the voting scheme have fewer false positives.

The slide in Figure 2.9 has G5 grade cancer which was not found in any other patients in the dataset. All networks had issues detecting cancer with this patient. There were many false negatives, regardless of if the networks were trained on raw images or TCMs. Xception-TCM nearly labelled all cancer as non-cancer. We saw this reflected in Figure 2.11 where FNR is shown to be is higher for G5 than G3 and G4 cancers for all models. This is likely because the network was not trained with sufficient samples to identify G5 cancer, even though the training data were balanced for cancer and non-cancer. If we increased the training dataset to include multiple patients with G5 cancer, then the network may have a lower FNR. Receiver operating characteristic curve (ROC) plots along with the AUC for this patient have been plotted in Figure 2.8.

Figure 2.10 has fewer false negatives than Figure 2.9. AlexNet-TCM has a large number of false negatives, which is not seen in other models. In this example, the location of false positives tend to be consistent across models. For example, along the edge of the bottom right of the gland, there tends to be a crescent of false positives seen across models and training data modalities. In this situation, some false positives are in the same area of prostatic intraepithelial neoplasia (PIN). PIN may be the confounding factor for this situation; however, this does not explain all cases of false positives.

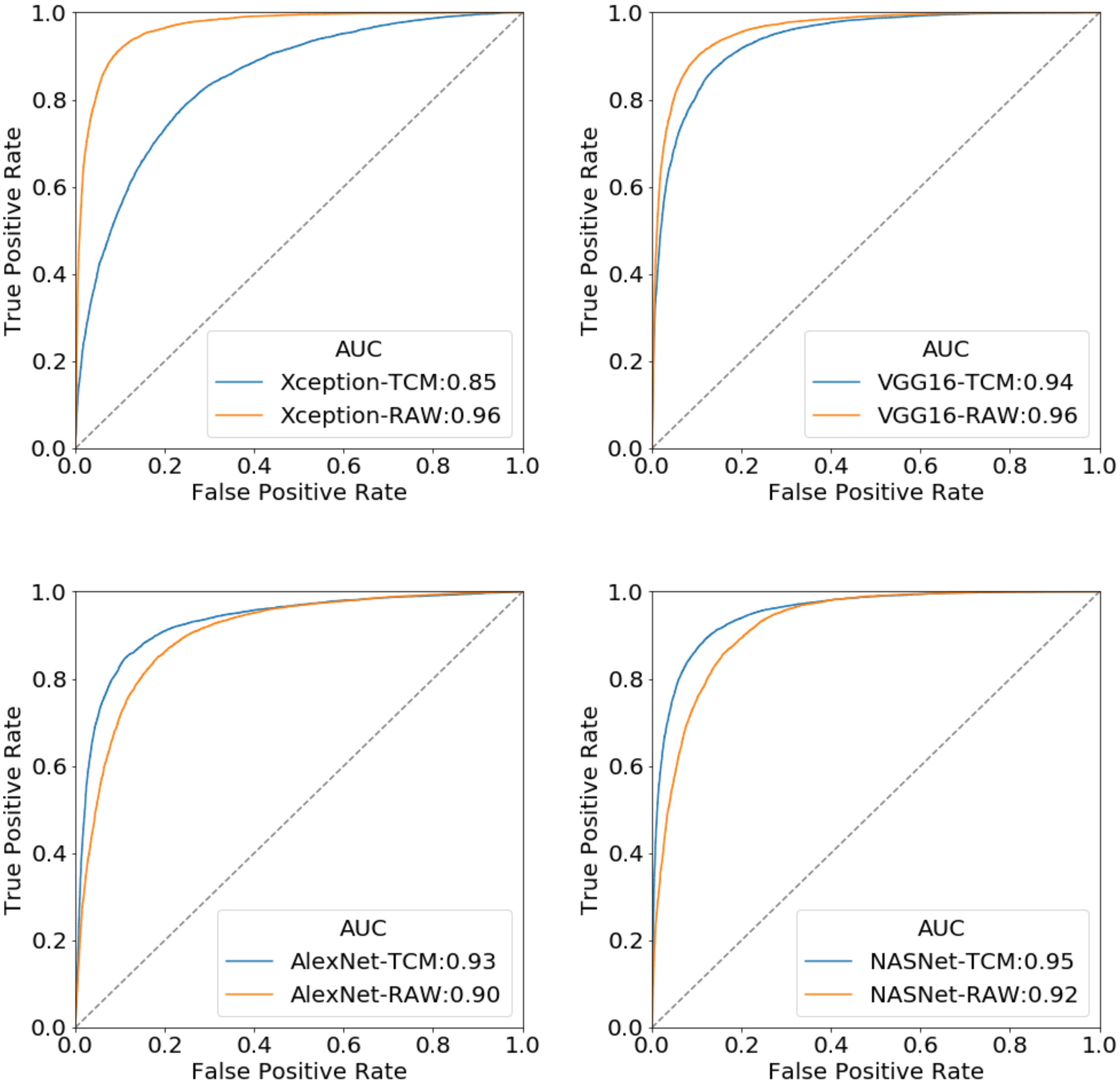
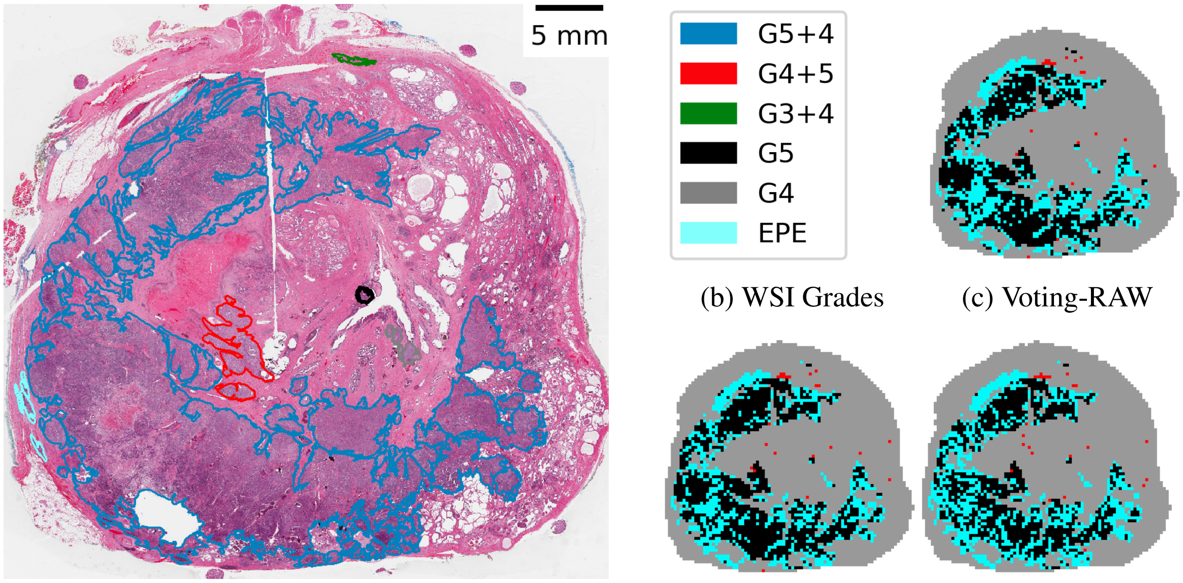
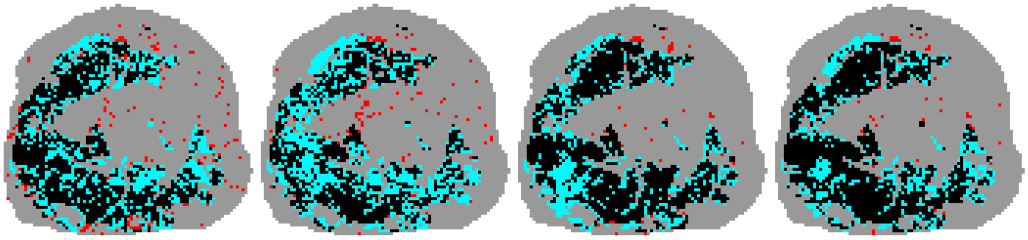


Figure 2.8: ROC plots with AUC for patient with G5 cancer in Figure.2.9

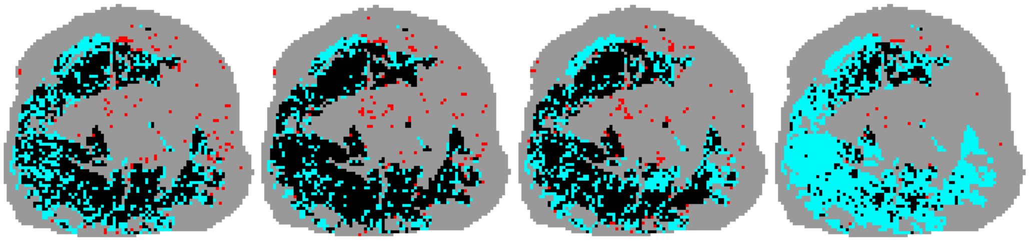


(b) WSI Grades (c) Voting-RAW

(a) WSI (d) Voting-All (e) Voting-TCM

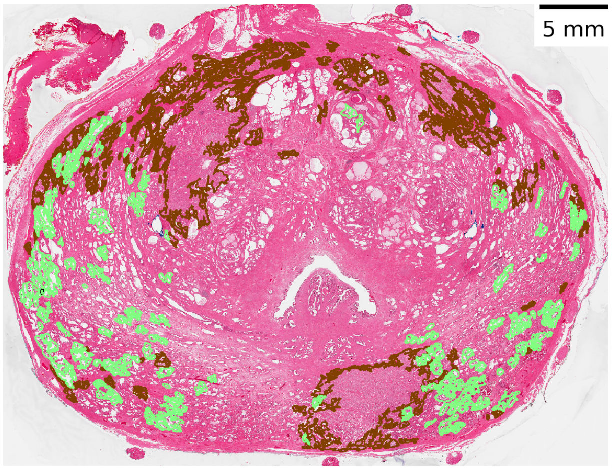


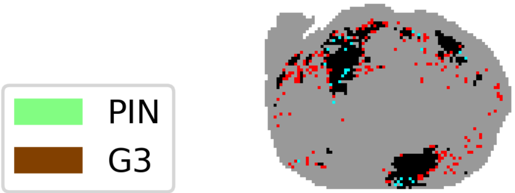
(f) AlexNet-RAW (g) NASNet-RAW (h) VGG16-RAW (i) Xception-RAW



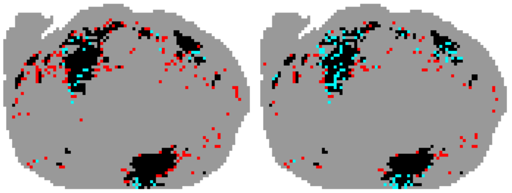
(j) AlexNet-TCM (k) NASNet-TCM (l) VGG16-TCM (m) Xception-TCM

Figure 2.9: Whole mount histology slide cancer maps. The cancer maps were produced from AlexNet, NASNet, VGG16, or Xception trained on either raw images or TCM. (a) is the WSI raw slide, (b) is the grade legend for the WSI and (c)-(m) are cancer maps where red is false positive, cyan is false negative, black is true positive and grey is true negative.

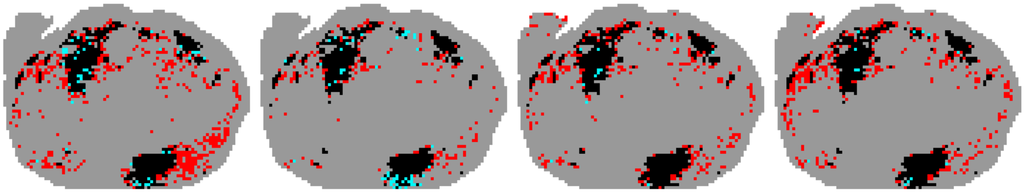




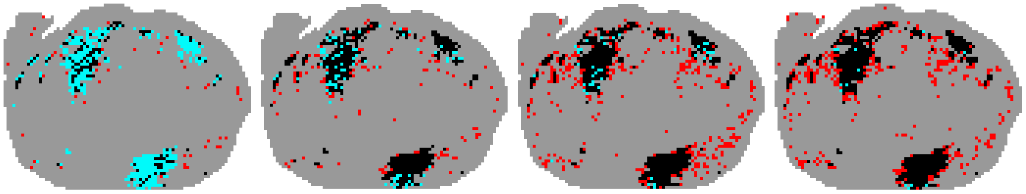
(b) WSI Grades (c) Voting-RAW



(a) WSI (d) Voting-All (e) Voting-TCM

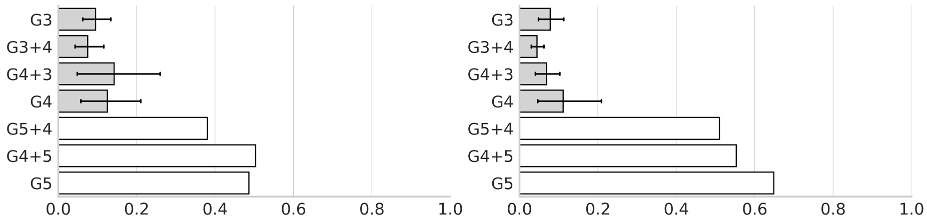


(f) AlexNet-RAW (g) NASNet-RAW (h) VGG16-RAW (i) Xception-RAW

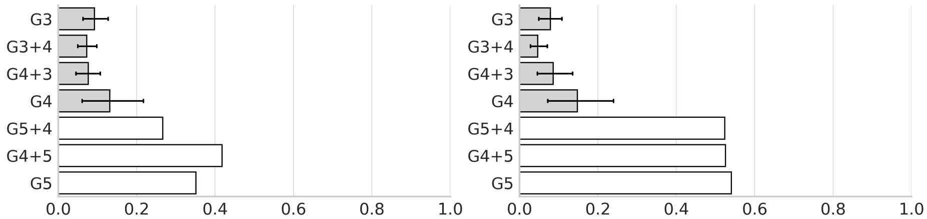


(j) AlexNet-TCM (k) NASNet-TCM (l) VGG16-TCM (m) Xception-TCM

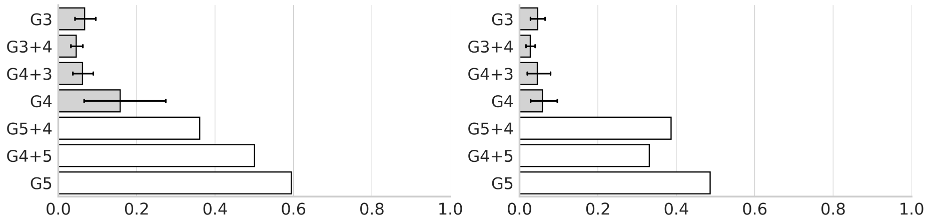
Figure 2.10: Whole mount histology slide cancer maps. The cancer maps were produced from AlexNet, NASNet, VGG16, or Xception trained on either raw images or TCM. (a) is the WSI raw slide, (b) is the grade legend for the WSI and (c)-(m) are cancer maps where red is false positive, cyan is false negative, black is true positive and grey is true negative.



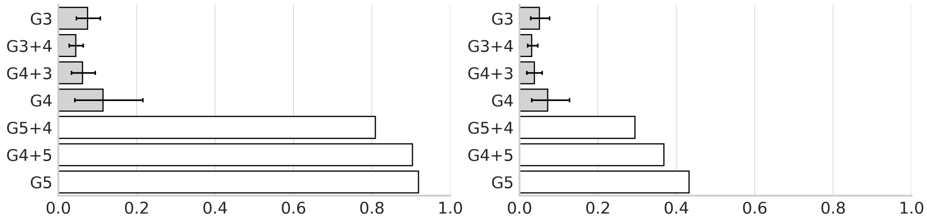
(a) AlexNet-TCM (b) AlexNet-RAW



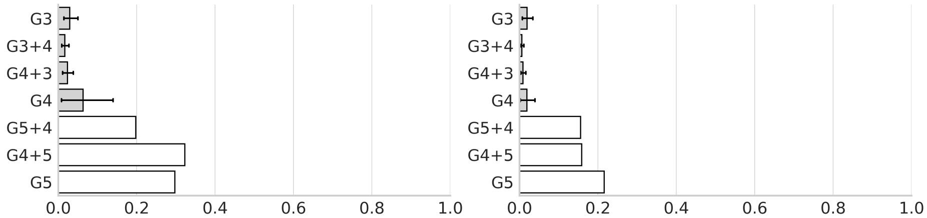
(c) NASNet-TCM (d) NASNet-RAW



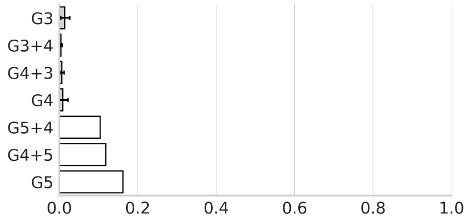
(e) VGG16-TCM (f) VGG16-RAW



(g) Xception-TCM (h) Xception-RAW



(i) Voting-TCM (j) Voting-RAW



(k) Voting-All

Figure 2.11: FNR of different cancer grades. Error bars indicate 95% CI. No error bars were given for G5, G5+4 and G4+5 because labels were taken from only one patient.

2.4.3 Limitations

The results of this study need to be considered in the context of its limitations. First, the U-Net generated TCMs were validated using segmentations generated by a single person, yielding the potential for bias toward under or over segmentation in the ground truth. Second, cancer classification was done using a LOPO cross validation scheme and was not validated on an external dataset; this will need to be performed in a future study as a critical step toward clinical translation. Third, our IGPC dataset had a very limited amount of G5 cancer, limiting our interpretation of the systems performance in classifying cancer of this grade.

2.5 Conclusions

In this work, we generated TCMs from H&E-stained radical prostatectomy specimens using U-Net. In contrast to previous work, U-Net does not require any thresholds to be set by the user. The system provided AUC values of greater than 0.96 for all tissue components, even when tested on images obtained from a different centre, demonstrating that the U-net could provide accurate TCMs for multiple centres without requiring retraining on data from each centre.

We tested the performance of multiple deep learning methods for classifying cancer vs. non-cancer using both TCMs and raw images. In general, models trained on raw images performed better than those trained on TCMs, although in many instances the differences were small and/or not statistically significant. That suggests the nuclei and lumina identified in the TCMs provide the majority of the information required for the models to accurately classify cancer vs. non-cancer. However, future work should include identifying the additional tissue components required for the models to equal the performance of classification using raw images. This yields important insight into the aspects of the tissue the models are using to make their classifications, potentially increasing pathologists confidence in the models and therefore supporting clinical translation.

Future work in this area includes user studies to assess whether cancer maps generated by neural networks will improve pathologist efficiency or accuracy, and whether including cancer maps in pathology reports will improve the decision to offer and target adjuvant therapy and ultimately—patient outcomes.

Bibliography

[1] F. Bray, J. Ferlay, I. Soerjomataram, *et al.*, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians* **68**(6), 394–424 (2018).

[2] I. M. van Oort, C. A. Hulsbergen-vandeKaa, and J. A. Witjes, "Prognostic Factors in Radical Prostatectomy Specimens: What Do We Need to Know from Pathologists?," **7**(12), 715–722 (2008).

[3] C. Magi-Galluzzi, A. J. Evans, B. Delahunt, *et al.*, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 3: Extraprostatic extension, lymphovascular invasion and locally advanced disease," *Modern Pathology* **24**(1), 26–38 (2011).

[4] P. H. Tan, L. Cheng, J. R. Srigley, *et al.*, "International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 5: Surgical margins," *Modern Pathology* **24**(1), 48–57 (2011).

[5] J. Croke, J. Maclean, B. Nyiri, *et al.*, "Proposal of a post-prostatectomy clinical target volume based on pre-operative MRI: volumetric and dosimetric comparison to the RTOG guidelines.," *Radiation Oncology* **9**, 303 (2014).

[6] S. J. Freedland, E. B. Humphreys, L. A. Mangold, *et al.*, "Risk of prostate cancer-specific mortality following biochemical recurrence after radical prostatectomy," *Journal of the American Medical Association* **294**(4), 433–439 (2005).

[7] T. A. Stamey, J. E. McNeal, C. M. Yemoto, *et al.*, “Biological determinants of cancer progression in men with prostate cancer,” *Journal of the American Medical Association* **281**(15), 1395–1400 (1999).

[8] J. I. Epstein, M. Carmichael, A. W. Partin, *et al.*, “Is tumor volume an independent predictor of progression following radical prostatectomy? A multivariate analysis of 185 clinical stage B adenocarcinomas of the prostate with 5 years of followup,” *Journal of Urology* **149**(6), 1478–1481 (1993).

[9] I. M. Thompson, R. K. Valicenti, P. Albertsen, *et al.*, “Adjuvant and salvage radiotherapy after prostatectomy: AUA/ASTRO guideline,” *Journal of Urology* **190**(2), 441–449 (2013).

[10] K. Moon, G. J. Stukenborg, J. Keim, *et al.*, “Cancer incidence after localized therapy for prostate cancer,” *Cancer* **107**(5), 991–998 (2006).

[11] K. Chrouser, B. Leibovich, E. Bergstralh, *et al.*, “Bladder cancer risk following primary and adjuvant external beam radiation for prostate cancer,” *Journal of Urology* **174**(1), 107–110 (2005).

[12] T. M. Pisansky, I. M. Thompson, R. K. Valicenti, *et al.*, “Adjuvant and Salvage Radiotherapy after Prostatectomy: ASTRO/AUA Guideline Amendment 2018-2019,” *The Journal of Urology* **202**(3), 533–538 (2019).

[13] T. Xia, A. Kumar, D. Feng, *et al.*, “Patch-level Tumor Classification in Digital Histopathology Images with Domain Adapted Deep Learning,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, **2018**, 644–647 (2018).

[14] W. Han, C. Johnson, M. Gaed, *et al.*, “Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens,” *Scientific Reports* **10**(1), 9911 (2020).

[15] U. A. H. Khan, C. Stürenberg, O. Gencoglu, *et al.*, "Improving Prostate Cancer Detection with Breast Histopathology Images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11435**, 91–99 (2019).

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, *et al.*, Eds., 1097–1105, Curran Associates, Inc. (2012).

[17] J. T. Kwak and S. M. Hewitt, "Nuclear Architecture Analysis of Prostate Cancer via Convolutional Neural Networks," *IEEE Access* **5**, 18526–18533 (2017).

[18] A. Madabhushi and M. Feldman, "Prostate Fused-MRI Pathology," *The Cancer Imaging Archive* (2016).

[19] K. Clark, B. Vendt, K. Smith, *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging* **26**(6), 1045–57 (2013).

[20] W. Han, A. D. Ward, C. Johnson, *et al.*, "Automatic cancer detection and localization on prostatectomy histopathology images ," in *Medical Imaging 2018: Digital Pathology*, J. E. Tomaszewski and M. N. Gurcan, Eds., **10581**, 205–212, International Society for Optics and Photonics, SPIE (2018).

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, **9351**, 234–241 (2015).

[22] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," *ICLR: International Conference on Learning Representations* (2015).

[23] B. Zoph, V. Vasudevan, J. Shlens, *et al.*, “Learning transferable architectures for scalable image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710 (2018).

[24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” (2014).

[25] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 1800–1807 (2017).

[26] J. Deng, W. Dong, R. Socher, *et al.*, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).

Chapter 3

General Discussion and Conclusions

3.1 Contributions of the Thesis

The thesis addresses questions proposed in Chapter 1. They are broken into two major questions outlined below.

(1) Can a U-Net trained to create tissue component maps at one centre, accurately segment slides at another? To answer this question, we trained U-Net to segment histology into nuclei and lumen maps, before validating on an external dataset. The trained U-Net performed similarly on both sets of data, showing that a U-Net can be used at a centre different than where it was trained. The AUC calculated using data from the training centre was 0.975 (95% confidence interval 0.969-0.980) for lumen and 0.983 (0.981-0.985) for nuclei. For the external dataset, the AUC was 0.980 (0.977-0.984) for lumina and 0.994 (0.993-0.996) for nuclei.

(2) How do DNNs compare when used for cancer vs. non-cancer classification in RP histology? We compared four deep learning models: AlexNet, NASNet, Xception and VGG16 trained on either raw images or TCMs. We found that AlexNet, NASNet and VGG16 trained on raw images perform with higher AUC and lower FPR and error rate, than the same model trained on TCMs. Xception models had no statistically significant difference between the two training image types in error rate and FPR. All models performed similarly and were trained and tested on the same dataset. The majority voting ensemble model had the lowest FPR and error rate—which is not surprising as outliers in one model would be removed if multiple models were used.

Every model had a higher FNR when classifying G5, G4+5 and G5+4 as cancer in comparison to G3, G3+4, G4+3 and G4. Our dataset did not have a balanced representation of all Gleason grading, where G5, G4+5 and G5+4 was only found in one patient. Consequently, this suggests that a balanced distribution of Gleason grades must be included in the training set and balancing for cancer vs non-cancer is not sufficient for cancer detection.

3.2 Limitations

We have not tested the cancer detection models on an external dataset, thus we do not have information on how the models will perform at multiple centres. This limited dataset and CV testing meant our data is subject to a positive bias. The U-Net was tested on external data, but this was only from one other centre. Staining variation arises from differences in tissue processing, but we have not tested how these models perform against a range of staining intensities. Therefore, our conclusions are limited by our datasets.

Our study used supervised learning. To label the data, cancer maps were generated by one physician and verified by one of two pathologists. Validation TCMs were generated by only one person. Our data are therefore biased towards the annotator. With 68 patients, this method of supervised learning using manual contours was feasible, but to scale this study to include hundreds or thousands of patients, it is impractical to expect detailed contours. This is why unsupervised or partially supervised models may be a more practical and scalable solution. Our method is limited in its ability to scale.

In the thesis, we covered several DNNs but there are many architectures that were not tested. Since we only sampled a small subset, this study can only reflect the performance of the four DNNs and cannot represent all DNNs. This study also only covered three component

TCMs at one resolution. It is possible that including certain components and excluding others may train for better networks, so we cannot draw conclusions on TCMs as a whole.

3.3 Applications and Future Directions

One of our goals from Chapter 2 was to automatically generate cancer maps, which may build towards our long-term goal of improving pathologist workflow. We analyzed FNR, FPR, error rate and AUC of the cancer maps, but we have not measured the impacts of automatically generated cancer maps on pathologist work-flow and on treatment decisions. Future research in this topic may allow us to better understand how we should optimize cancer map generation and if these cancer maps improve treatment outcomes.

To improve pathologists' workflow, their perceptions of a DNN assistive program and how the assistance should be provided, is important to research. A study surveying pathologists' perspectives of artificial intelligence (AI) in pathology found 48% of respondents felt that diagnostic decisions should remain predominantly human, 25% found it should be an equal role and 20% found AI should take a dominant role [1]. This brings up the question of how much control a DNN assistive program should have. Campenella et al. argues that in a clinical setting, difficult cases are often reviewed by multiple pathologists and it can be assumed that a comprehensive centre has 100% accuracy, thus assistive algorithms that filter suspicious results for a pathologist only need to have 100% sensitivity and an acceptable FPR [2]. Their work suggests DNNs should act like a filtering program with the final decision made by the pathologist, which is supported by the aforementioned survey. Initial work has been done studying AI assistive programs effect on sensitivity and observer agreement in PCa biopsy, but to the best of my knowledge this has not been done for RP [3, 4]. Therefore, it is important to implement a user study to understand how machine learning (ML) can be incorporated and how pathologists should interface with a DNN assistive program in RP histology.

The International Society of Urological Pathologist (ISUP) concluded it is important to provide information on where EPE is located and the location of positive margins [5, 6]. But, we do not know if patient specific cancer maps, rather than a written description of where the cancer is located, would better predict post surgery outcome or guide targeted adjuvant radiotherapy to the prostate bed. Consequently, studying the impact of cancer maps on treatment outcomes is a potential research project.

ML, in general, performs poorly on outliers. We saw this in Chapter 2 where G5 grade ROIs (outliers in our dataset) had a larger FNR than G4 and G3 even when the problem was cancer vs. non-cancer classification rather than Gleason grading. In the future, studies involving a balanced set of Gleason scores may solve that issue but for other rare situations, balancing may not be possible. Rare types of cancer, like large cell prostate cancer, are incredibly aggressive and may be missed when using ML algorithms, leading to poor patient care. It is often difficult to collect data on those types of cancer due to the rarity, so pipelines involving ML algorithms should take these rare cases into consideration.

We also can consider ethics and legal issues surrounding machine learning in a clinical setting. Schiff et al. outlines the issue of informed consent [7] relative to ML. They ask—to what level of understanding should a physician have of ML theory to adequately inform a patient for informed consent? Given the black box nature of ML and the poor performance on outliers (like rare types of cancer), can ML models guarantee a 100% sensitivity suggested by Campenella et al. [2]? What is an acceptable failure rate and who would be responsible legally and ethically when/if it does? Sullivan et al. concludes that current legal models do not adequately address situations in which AI results in injury [8]. Other concerns include economic displacement due to AI tools where 18% of surveyed pathologists were concerned and 2% were extremely concerned about job displacement, and 26% expressed concern that AI would erode pathologists' skills [1]. Ethical questions surrounding ML in medicine are difficult to answer, but must be solved before clinical applications are made.

Some of these questions are not new and are required when any new technology is applied to the medical field. Still, these questions surrounding the ethics of AI in healthcare are difficult to answer, but should be addressed before clinical implementation.

Bibliography

[1] S. Sarwar, A. Dent, K. Faust, *et al.*, “Physician perspectives on integration of artificial intelligence into diagnostic pathology,” *NPJ Digital Medicine* **2**(1), 28 (2019).

[2] G. Campanella, M. G. Hanna, L. Geneslaw, *et al.*, “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nature Medicine* **25**(8), 1301–1309 (2019).

[3] P. Raciti, J. Sue, R. Ceballos, *et al.*, “Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies,” *Modern Pathology* (2020).

[4] W. Bulten, M. Balkenhol, J. J. A. Belinga, *et al.*, “Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists,” *Modern Pathology* (2020).

[5] C. Magi-Galluzzi, A. J. Evans, B. Delahunt, *et al.*, “International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. working group 3: Extraprostatic extension, lymphovascular invasion and locally advanced disease,” *Modern Pathology* **24**(1), 26–38 (2011).

[6] P. H. Tan, L. Cheng, J. R. Srigley, *et al.*, “International society of urological pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 5: Surgical margins,” *Modern Pathology* **24**(1), 48–57 (2011).

[7] D. Schiff and J. Borenstein, “How should clinicians communicate with patients about the roles of artificially intelligent team members?,” *AMA Journal of Ethics* **21**(2), 138–145 (2019).

[8] H. R. Sullivan and S. J. Schweikart, “Are current tort liability doctrines adequate for addressing injury caused by AI?,” *AMA Journal of Ethics* **21**(2), 160–166 (2019).

Chapter 4

Table of Abbreviations and Symbols

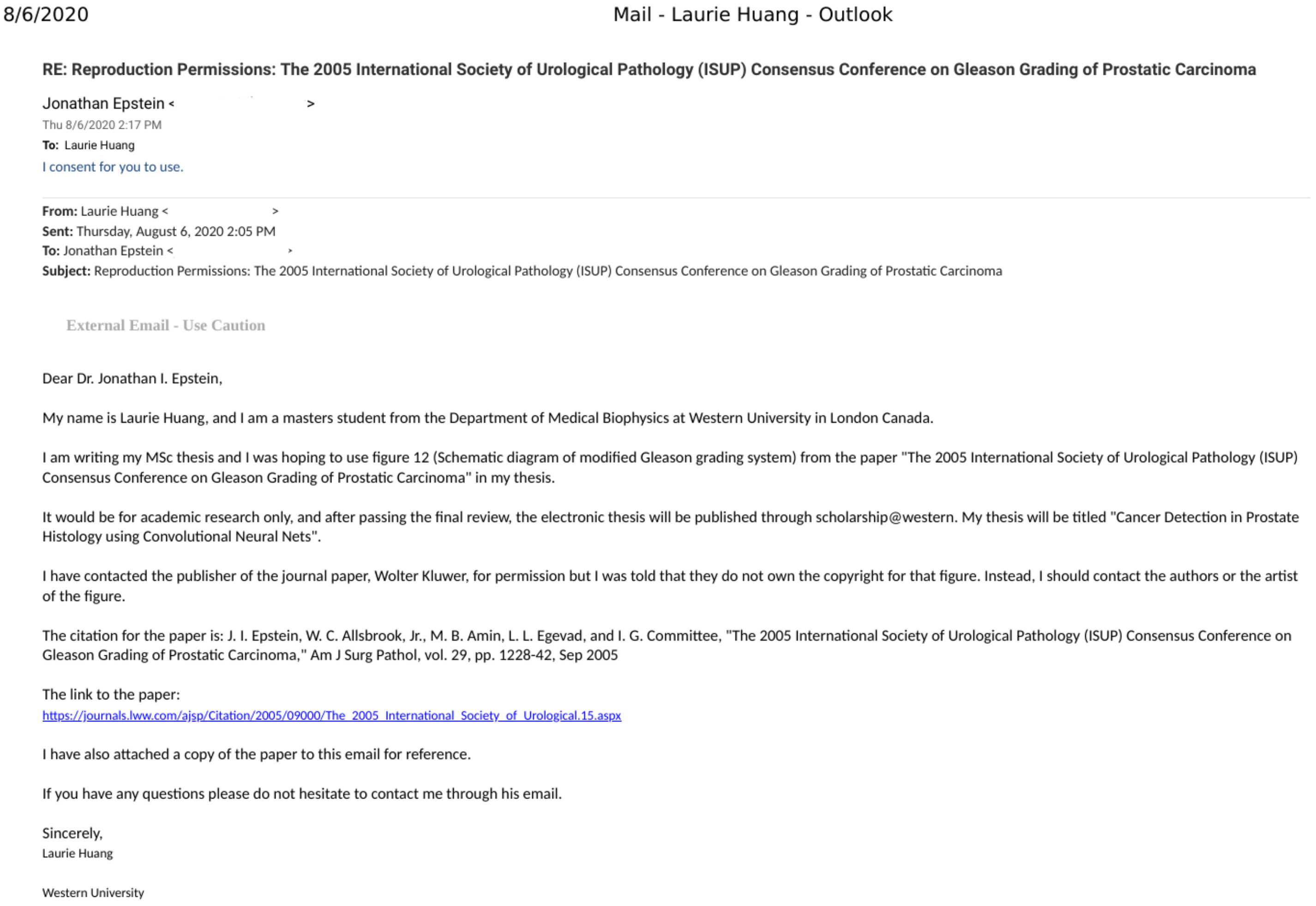
AI	Artificial Intelligence
ASTRO	American Society for Radiation Oncology
AUA	American Urological Society
AUC	Area Under the Receiver Operating Characteristic Curve
CI	Confidence Interval
CNN	Convolutional Neural Network
CV	Cross Validation
DL	Deep Learning
DNN	Deep Neural Network
EPE	Extraprostatic Extension
FNR	False Positive Rate
FP	False Positive
FPR	False Negative Rate
GS	Gleason Score
GPU	Graphics Processing Unit
H&E	Hematoxylin and Eosin
IGPC	Image Guidance for Prostate Cancer

ISBI	International Symposium on Biomedical Imaging
ISUP	The International Society of Urological Pathology
LOPO	Leave-One-Patient-Out
ML	Machine Learning
NN	Neural Network
PCa	Prostate Cancer
PIN	Prostatic Intraepithelial Neoplasia
RBCs	Red Blood Cells
ROC	Receiver Operator Curve
ROIs	Region of Interest
RP	Radical Prostatectomy
TCIA	The Cancer Imaging Archive
TCM	Tissue Component Map
TMA	Tissue Micro Arrays
WSI	Whole-Slide Images

Appendix A

Permissions to Reproduce Previously Published Material

Permission to reproduce published material in Figure 1.3 from Chapter 1.



8/6/2020 Mail - Laurie Huang - Outlook

RE: Reproduction Permissions: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma

Jonathan Epstein < >
Thu 8/6/2020 2:17 PM
To: Laurie Huang

I consent for you to use.

From: Laurie Huang < >
Sent: Thursday, August 6, 2020 2:05 PM
To: Jonathan Epstein < >
Subject: Reproduction Permissions: The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma

External Email - Use Caution

Dear Dr. Jonathan I. Epstein,

My name is Laurie Huang, and I am a masters student from the Department of Medical Biophysics at Western University in London Canada.

I am writing my MSc thesis and I was hoping to use figure 12 (Schematic diagram of modified Gleason grading system) from the paper "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma" in my thesis.

It would be for academic research only, and after passing the final review, the electronic thesis will be published through scholarship@western. My thesis will be titled "Cancer Detection in Prostate Histology using Convolutional Neural Nets".

I have contacted the publisher of the journal paper, Wolter Kluwer, for permission but I was told that they do not own the copyright for that figure. Instead, I should contact the authors or the artist of the figure.

The citation for the paper is: J. I. Epstein, W. C. Allsbrook, Jr., M. B. Amin, L. L. Egevad, and I. G. Committee, "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," Am J Surg Pathol, vol. 29, pp. 1228-42, Sep 2005

The link to the paper:
https://journals.lww.com/ajsp/Citation/2005/09000/The_2005_International_Society_of_Urological.15.aspx

I have also attached a copy of the paper to this email for reference.

If you have any questions please do not hesitate to contact me through his email.

Sincerely,
Laurie Huang

Western University

Curriculum Vitae

Laurie Huang

Education

2018-Present	M.Sc. in Medical Biophysics • Supervisor: Aaron Ward	*Western University*, London
2014-2018	B.Sc. in Physics (Honours)	*Queen's University*, Kingston

Awards

2017	Harold M. Cave Undergraduate Travel Scholarship	*Queen's University*, Kingston
2017	ORA Student Exchange Program Scholarship	*Ontario Universities International*, Toronto
2016	Undergraduate Student Research Award	*NSERC*, Canada
2016	Harold M. Cave Undergraduate Travel Scholarship	*Queen's University*, Kingston
2015	Dean's Honour List	*Queen's University*, Kingston
2014	Queen's University Excellence Scholarship	*Queen's University*, Kingston

Teaching Experience

2019	Teaching Assistant •CS2211a - Software Tools and Systems Programming	*Western University*, London
2019	Teaching Assistant Training Program	*Western University*, London
2017	Teaching Assistant •PHYS250 - General Laboratory	*Queen's University*, Kingston
2018	Teaching Assistant •PHYS372 - Thermodynamics	*Queen's University*, Kingston

Conferences

2020	Imaging Network Ontario Symposium L. Huang, W. Han, J.A. Gomez, M. Moussa, S.E. Pautler, J.L. Chin, G.S. Bauman, and A.D. Ward •*Tissue component segmentation and cancer detection on digitized prostatectomy slides using convolutional neural networks*	Toronto

Research

2017	Research Assistant	*Institut de planétologie et d'astrophysique de Grenoble*, Grenoble
2016	Research Assistant	*Royal Military College of Canada, Physics Department*, Kingston

Extracurricular Roles

2016-2018	CCUWIP Conference Organizer	*Queen's University*, Kingston
2015-2018	Queen's Brazilian Jujitsu Trainee	*Queen's University*, Kingston
2014-2018	Queen's Dance Club Trainee	*Queen's University*, Kingston